

# INFERENCE IN ADDITIVELY SEPARABLE MODELS WITH A HIGH DIMENSIONAL SET OF CONDITIONING VARIABLES

DAMIAN KOZBUR

**ABSTRACT.** This paper considers estimation and inference of nonparametric conditional expectation relations with a high dimensional conditioning set. Rates of convergence and asymptotic normality are derived for series estimators for models where conditioning information enters in an additively separable manner and satisfies sparsity assumptions. Conditioning information is selected through a model selection procedure which chooses relevant variables in a manner that generalizes the post-double selection procedure proposed in [9] to the nonparametric setting. The proposed method formalizes considerations for trading off estimation precision with omitted variables bias in a nonparametric setting. Simulation results demonstrate that the proposed estimator performs favorably in terms of size of tests and risk properties relative to other estimation strategies.

*Key Words:* nonparametric models, high dimensional-sparse regression, inference under imperfect model selection. *JEL Codes:* C1.

## 1. INTRODUCTION

Nonparametric estimation in economic and statistical problems is common because it is appealing in applications for which functional forms are unavailable. In many problems, the primary quantities of interest can be computed from the conditional expectation function of an outcome variable  $y$  given a regressor of interest  $x$ . Nonparametric methods are often attractive for estimating such conditional expectations since assuming an incorrect simple parametric model between the variables of interest will lead to incorrect inference.

In many econometric models, it is important to take into account conditioning information,  $z$ . When  $x$  is not randomly assigned, estimates of

---

*Date:* First version: September 2013. This version is of October 1, 2015.

*Correspondence:* Haldeneggsteig 4, 8092 Zürich, Center for Law and Economics, D-GESS, ETH Zürich, damian.kozbur@gess.ethz.ch.

I thank Christian Hansen, Tim Conley, Matt Taddy, Azeem Shaikh, Dan Nguyen, Emily Oster, Martin Schonger, Eric Floyd, and seminar participants at University of Western Ontario, University of Pennsylvania, Rutgers University, Monash University, Center for Law and Economics at ETH Zurich for helpful comments. I gratefully acknowledge financial support from the ETH Postdoctoral Fellowship .

partial effects of  $x$  on  $y$  will often be incorrect if  $z$  is ignored and at the same time, partly influences both  $x$  and  $y$ . However, if  $x$  can be considered as approximately randomly assigned given  $z$ , then the conditional expectation of  $y$  given  $x$  and  $z$  can be used to calculate causal effects of  $x$  on  $y$  and to evaluate counterfactuals. Therefore, properly accounting for conditioning information  $z$  is of primary importance.

When conditioning information is important to the problem, it is necessary to replace the simple objective of learning the conditional mean function  $E[y|x] = g(x)$  with the new objective of learning a family of conditional mean functions

$$E[y|x, z] = g_z(x) \quad (1.1)$$

indexed by  $z$ . Properly accounting for conditioning information in  $z$  may be done in several ways and leaves the researcher with important modeling decisions. For the sake of illustration, four potential ways to account for conditioning information are (1) by specifying a partially linear model  $g_z(x) = g(x) + z'\beta$ , (2) by specifying an additive model  $g_z(x) = g(x) + h(z)$ , (3) by specifying a multiplicative model  $g_z(x) = g(x)h(z)$ , or (4) by specifying a fully nonparametric model  $g_z(x) = g(x, z)$ . The fully nonparametric model suffers from the curse of dimensionality even for moderately many covariates, while the partially linear model may be too rigid and may miss important conditioning information.

Many applications have potentially large conditioning sets. A large conditioning set in economics may arise because the researcher wishes to control for many measured characteristics, like demographics, at the observation level. In the extreme cases, the dimension of  $z$  can be large enough so that all four example specifications for  $g_z(x)$  listed above require an infeasible amount of data in order to avoid statistical overfitting and ensure good inference.

This paper is restricted to studying the partially linear model and the additive model shown above.<sup>1</sup> Therefore, the interest is the specialization of model (1.1) to the case

$$E[y|x, z] = g(x) + h(z). \quad (1.2)$$

When additive the model provides a good approximation to the underlying data generating structure, it is useful since many quantities describing the relationship of  $x$  and  $y$  conditional on  $z$  can be learned with a good understanding of  $g(x)$  alone. In addition, it provides a clear description of how the conditional relation between  $x$  and  $y$  changes as  $z$  changes.

An important structure that has been used in recent econometrics is approximate sparsity. See for example [5], [3], and [9]. In the context of this paper, approximate sparsity informally refers to the condition that the conditional expectation function  $g_z(x)$  can be approximated by a family of

---

<sup>1</sup>These models have a particular structure which make them convenient for the problem of selecting a conditioning set. The two alternative models are likely to require conditions considerably different than the ones considered here.

functions which depend only on a small (though a priori unknown) subset of the conditioning information contained in  $z$ . Sparsity is useful because in principal, a researcher can address statistical overfitting problems by locating and controlling for only correct conditioning information. The focus of this paper is on providing a formal model selection technique over models for the conditioning set  $z$  which retrieve the relevant conditioning information. The retrieval of relevant conditioning information will be done in such a way so that standard estimation techniques performed after model selection will provide correct inference for nonparametric partial effects of  $x$ .

This paper takes model (1.2) as a starting point and assumes that  $h(z)$ , a complicated function of many conditioning variables, has sparse structure. The strategy for identifying a sparse structure is to search for a small subset of relevant terms within a long series expansion for  $h(z)$ . To formalize this, a high dimensional framework, allowing the long series expansion for  $h(z)$  to have more terms than the sample size is particularly convenient. Once a simple model for  $h(z)$  is found, the focus returns to the estimation of  $g(x)$ . The paper contributes to the nonparametrics literature by establishing rates of convergence of estimates and asymptotic normality for functionals of  $g(x)$  after formal model selection has been performed to simplify the way the conditioning variable  $z$  affects the conditional expectation relation between  $x$  and  $y$ .

In addition to addressing questions about flexible selection of conditioning set in a nonparametric setting, this paper contributes to a broader program aimed at conducting inference in the context of high-dimensional models. Statistical methods in high dimensions have been well developed for the purpose of prediction. Two widely used methods for estimating high dimensional predictive relationships and are important for the present paper are Lasso and Post Lasso. The Lasso is a shrinkage procedure which estimates regression coefficients by minimizing a loss function plus a penalty for the size of the coefficient. Post-Lasso fits an ordinary least squares regression on variables with non-identically-zero estimated Lasso coefficients. For theoretical and simulation results about the performance of these two methods, see [18] [37], [19] [15] [1], [2], [10], [13], [12] [14], [15], [20], [23], [24], [26], [27], [28], [33], [37], [38], [40], [41], [4], [11], [4], among many more. Regularized estimation buys stability through reduction in estimate variability at the cost of a modest bias in estimates. Regularized estimators like the Lasso where many parameter values are set identically to zero, also favor parsimony. Recently, several authors have begun the task of assessing uncertainties or estimation error of model parameter estimates in a wide variety of models with high dimensional regressors (see, for example, [5]; [3]; [42]; [6]; [9]; [39]; [21]; and [8]).

Quantifying estimation precision has been shown to be difficult theoretically (for formal statements, see [31], [25]) because model selection mistakes and regularization typically bias estimates to the same order of magnitude (relative to the sample size) as estimation variability. This paper builds on

methodology found in [9] (named Post-Double-Selection) which gives robust statistical inference for the slope parameter of a treatment variable  $x$  with high-dimensional confounders  $z$  in the context of a partially linear model  $E[y|x, z] = \alpha x + z'\eta$ .<sup>2</sup> The method selects elements of  $z$  in two steps: step 1 selects the terms in  $z$  that are most useful for predicting  $x$ , and step 2 selects elements of  $z$  most useful for predicting  $y$  in a second step. The use of two model selection steps is motivated partially by the intuition that two necessary conditions for omitted variables bias to occur: an omitted variable exists which is (1) correlated with the treatment  $x$ , and (2) correlated with the outcome  $y$ . Each selection step addresses one of the two concerns. In their paper, they prove that under the regularity right conditions, the two described model selection steps can be used to obtain asymptotically normal estimates of  $\alpha$  and in turn to construct correctly sized confidence intervals. This paper generalizes the approach from estimating a linear treatment model to estimating a component in nonparametric additively separable models. The main technical contribution is providing conditions under which nonparametric estimates for functionals of  $g(x)$  are uniformly asymptotically normal after model selection on a conditioning set over a large set of data generating processes.

## 2. A HIGH DIMENSIONAL ADDITIVELY SEPARABLE MODEL

This section provides an intuitive discussion of the additively separable nonparametric model explored in this paper. Recall the additive conditional expectation model described in the introduction:

$$E[y|x, z] = g_z(x) = g(x) + h(z).$$

The interest is in recovering the function  $g(x)$  which describes the conditional relationship between the treatment variable of interest,  $x$ , and the outcome  $y$ . The component functions  $g$  and  $h$  belong to ambient spaces  $g \in \mathcal{G}, h \in \mathcal{H}$  which restricted sufficiently to allow  $f$  and  $g$  to be uniquely identified.<sup>3</sup> The function  $h$  and the variable  $z$  will be allowed to depend on  $n$  to facilitate a high-dimensional thought experiment for the conditioning set. As an example, this allows estimation of models of the form  $E[y|x, z] = g(x) + z'_n \beta_n$  with  $\dim(z_n) \rightarrow \infty$ . Dependence on  $n$  will be suppressed for ease of notation. The formulation will allow  $x$  and  $z$  to share variables to some extent. For instance, the setup will allow for additive interaction models like those found in Andrews and Whang (1991) so that the model  $E[y|x, z] = g(x) + \gamma \cdot x \cdot z + h(z)$  where  $\gamma$  is an unknown scalar.

The estimation of  $(g, h)$  proceeds by a series approximation with a dictionary that is partially data-dependent. As a review, a standard series estimator of the conditional expectation function without conditioning set,

---

<sup>2</sup>Their formulation is slightly more general because it allows the equality to be an approximate equality.

<sup>3</sup>For example, it necessary to require additional conditions like  $g(0) = 0$ , otherwise, at best,  $g$  and  $h$  are identified up to addition by constants

$g(x) = E[y|x]$ , is obtained with the aid of a dictionary of transformations  $p^K(x) = (p_{1K}(x), \dots, p_{KK}(x))'$ . The dictionary consists of a set of  $K$  functions of  $x$  with the property that a linear combination of the  $p_{jK}(x)$  can approximate  $g$  to an increasing level of precision that depends on  $K$ .  $K$  is permitted to depend on  $n$  and  $p^K(x)$  may include splines, fourier series, orthogonal polynomials or other functions which may be useful for approximating  $g$ . The series estimator is simple and implemented with standard least squares regression. Given data  $(y_i, x_i)$  for  $i = 1, \dots, n$ , a series estimator for  $g$  is takes the form:  $\hat{g}(x) = p^K(x)' \hat{\beta}$  for  $P = [p^K(x), \dots, p^K(x_n)]'$ ,  $Y = (y_1, \dots, y_n)'$  and  $\hat{\beta} = (P'P)^{-1}P'Y$ . Traditionally, the number of series terms, chosen in a way to simultaneously reduce bias and increase precision, must be small relative to the sample size. Thus the function of interest must be sufficiently smooth or simple in order for nonparametric estimation to work well. The econometric theory for nonparametric regression estimation using an approximating series expansion is well-understood under standard regularity conditions; see, for example, [36], [16], [29].

Series estimation is particularly convenient for estimating models of the form of (1.2) because they can be approached using two dictionaries,  $p^K(x), q^L(z)$  consisting of  $K$  and  $L$  terms which individually approximate  $g(x)$  and  $h(z)$ . The dictionaries can simply be combined into one larger dictionary. To describe the estimation procedure in this paper, suppose such a dictionary,  $(p^K(x), q^L(z)) = (p_{1K}(x), \dots, p_{KK}(x), q_{1L}(z), \dots, q_{LL}(z))$ , compatible with the additively separable decomposition exists and is known. In what follows, dependence on  $K$  and  $L$  is suppressed in the notation so that  $p^K(x) = p(x)$  and  $q^L(z) = q(z)$ .

The two dictionaries differ in nature. The first dictionary,  $p(x)$  is traditional, and follows standard conditions imposed on series estimators, for example, [29], requiring among other conditions, that  $K \rightarrow \infty, K/n \rightarrow 0$ . The first dictionary must be chosen to approximate the function  $g(x)$  sufficiently well so that if  $h(z)$  were known exactly,  $g(x)$  could be estimated in the traditional nonparametric way and inferences on functionals of  $g(x)$  would be reliable.

The second dictionary,  $q(z)$ , is afforded much more flexibility. This is convenient and appropriate when entertaining a high dimensional conditioning set  $z$ . When the problem of interest is in recovering and performing inference for  $g(x)$ , the second component  $h(z)$  may be considered a nuisance parameter. In particular, this paper will not be concerned with constructing confidence intervals for  $h(z)$ , and therefore the requirements on the magnitude of bias in estimating  $h(z)$  will be less stringent. As a consequence, model selection bias of estimates of  $h(z)$ , when done according to the method below, will have negligible impact on the coverage probabilities of confidence sets. Increased flexibility in modeling  $h(z)$  by allowing  $L > n$  can make subsequent inference for  $g(x)$  more robust, but requires additional structure on  $q(z)$ . The key additional conditions are sparse approximation conditions.

The first sparsity requirement is that there is a small number of components of  $q(z)$  that adequately approximate the function  $h(z)$ . The second sparsity requirement is that information about functions  $h \in \mathcal{H}$  conditional on  $\mathcal{G}$  can be suitably approximated using a small number of terms in  $q(z)$ . The identities of the contributing terms, however, can be unknown to the researcher a priori.

Aside from estimating an entire component of conditional expectation function  $g(x)$  itself, a goal of this paper is to obtain asymptotically normal estimates of certain functionals of  $g(x)$ . Given a functional  $a(g)$ , that satisfies certain regularity conditions, the model selection procedure on  $q(z)$  will deliver a model such that subsequent plug-in estimates  $a(\hat{g})$  will be asymptotically normal around  $a(g)$ . Such functionals include integrals of  $g$ , weighted average derivatives of  $g(x)$ , evaluation of  $g(x)$  at a point  $x^0$ , and  $\arg \max g(x)$ .

### 3. ESTIMATION

When the number of free parameters is larger than the sample size, model selection or regularization is necessary. There are a variety of different model selection techniques available to researchers. A popular approach is via the Lasso estimator given by [18] and [37] which in the context of regression, simultaneously performs regularization and model selection. The Lasso is used in many areas of science and image processing and has demonstrated good predictive performance. Lasso allows the estimation of regression coefficients even when the sample size is smaller than the number of parameters by adding to the quadratic objective function a penalty term which mechanically favors regression coefficients that contain zero elements. By taking advantage of ideas in regularized regression, this paper demonstrates that quality estimation of  $g(x)$  can be attained even when  $K + L$ , the effective number of parameters, exceeds the sample size  $n$ . Estimating proceeds by a model selection step that effectively reduces the number of parameters to be estimated. There are many other sensible candidates for model selection devices in the statistics and econometrics literature. The appropriate choice of model selection methodology can be tailored to the application. In addition to the Lasso, variants of Lasso like the group-Lasso, the Scad (see [17]), the BIC, the AIC all feasible examples. In the exposition of the results, the model selection procedure used will be specifically the Lasso because it is simple and widely used. The section 3.2 below provides a brief review of Lasso, especially those that arise in econometric applications.

Estimation of  $E[y|x, z]$  will be based on a reduced dictionary  $(\tilde{p}(x), \tilde{q}(z))$  comprised of a subset of the series terms in  $p(x)$  and  $q(x)$ . Because the primary object of interest is  $g(x)$ , it is natural to include all terms belonging to  $p(x)$  in the reduced dictionary, giving  $\tilde{p}(x) \equiv p(x)$ . Therefore, the main selection step involves choosing a subset of terms from  $q(z)$ . Given a model selection procedure which provides a new reduced dictionary  $\tilde{q}(z)$ , containing

$\tilde{L} < L$  series terms, the post-model selection estimate of  $g(x)$  is defined by

$$\hat{g}(x) = p(x)' \hat{\beta}$$

where  $[\hat{\beta}' \ \hat{\eta}']' := ([P \ \tilde{Q}]' [P \ \tilde{Q}])^{-1} [P \ \tilde{Q}]' Y$ .

Since estimation of  $h(z)$  is of secondary concern, only the components of  $h(z)$  predictive of  $g(x)$  and  $y$  need to be estimated. These two predictive goals will guide the choice of model selection procedure as described in the upcoming sections. The results will demonstrate that under standard regularity, the post-model selection estimates give convergence rates for  $\hat{g}(x)$  which are the same as in classic nonparametric estimation as well as asymptotic normality results for plug in estimators,  $\hat{a}(g) = a(\hat{g})$  of nonlinear functionals  $a$  of the underlying conditional expectation function.

### 3.1. Nonparametric Post-Double Selection in the Additive Model.

The main challenge in statistical inference after model selection is in attaining robustness to model selection errors. When coefficients are small relative to the sample size (ie statistically indistinguishable from zero), model selection mistakes are unavoidable.<sup>4</sup> When such errors are not accounted for, subsequent inference has been shown to be potentially severely misleading. This intuition is formally developed in [31] and [25]. Offering solutions to this problem is the focus of a number of recent papers; see, for example, [5], [3], [42], [6], [9], [39], [21], [8], and [7].<sup>5</sup> This section extends the approach of [9] to the nonparametric setting.

Informally, model selection in the additively separable model proceeds in two steps. The two selection steps are based on the observation that the functional relation  $g(x)$  can be learned with knowledge of the conditional expectations

$$E[\varphi(x)|z] \tag{3.3}$$

$$E[y|z] \tag{3.4}$$

for a robust enough family of test functions  $\varphi(x)$ , for instance, smooth functions with compact support. Equivalently, the relation  $g(x)$  can be learned by projecting out the variable  $z$  and working with residuals. In the additively separable model, the two selection steps are summarized as follows:

(1) *First Stage Model Selection Step* - Select those terms in  $q$  which are relevant for predicting terms in  $p$ .

(2) *Reduced Form Model Selection Step* - Select those terms in  $q$  which are relevant for predicting  $y$ .

---

<sup>4</sup>Under some restrictive conditions, for example beta-min conditions which constrain nonzero coefficients to have large magnitudes, perfect model selection can be attained.

<sup>5</sup>Citations are ordered by date of first appearance on arXiv.



To further describe the selection stages, it is convenient to ease notation by introducing an operator  $T$  on functions that belong to  $\mathcal{G}$ :

$$T\varphi(z) = E[\varphi(x)|z]$$

This notion is convenient for understanding the validity behind post double selection in the additively separable model. The operator  $T$  measures dependence between functions in the ambient spaces  $\mathcal{G}, \mathcal{H}$  which house the functions  $g, h$  and the conditioning is understood to be on all function  $\varphi \in \mathcal{H}$ .

If the operator  $T$  can be suitably well approximated, then the post double selection methodology generalizes to the nonparametric additively separable case. The operator  $T$  on  $\phi$  will be approximated as a linear combination, given by  $\Gamma_\phi$  of basis terms  $q$  so that

$$T\varphi(z) \approx q(z)' \Gamma_\varphi.$$

Meanwhile,  $\Gamma_\varphi$  is approximated with linear combinations of  $\Gamma_{p_k}$ ,  $1 \leq k \leq K$ . The final selected model  $\tilde{q}$  consists of the union of terms selected during the first stage model selection step and the reduced form model selection step. A practical implementation algorithm is provided in Section 3.3

[9] develop and discuss the post-double-selection method in detail for partially linear model. They note that including the union of the variables selected in each variable selection step helps address the issue that model selection is inherently prone to errors unless stringent assumptions are made. As noted by [25], the possibility of model selection mistakes precludes the possibility of valid post-model-selection inference based on a single Lasso regression within a large class of interesting models. The chief difficulty arises with covariates whose effects in (3.3) are small enough that the variables are likely to be missed if only (3.3) is considered but have large effects in (3.4). The exclusion of such variables may lead to substantial omitted variables bias if they are excluded which is likely if variables are selected using only (3.3).<sup>6</sup> Using both model selection steps guards against such model selection mistakes and guarantees that the variables excluded in both model selection steps have a negligible contribution to omitted variables bias under the conditions listed below.

**3.2. Brief overview of Lasso methods.** The following description of the Lasso estimator is a review of the particular implementation given in [3]. Consider the conditional expectation  $E[y|w] = f(w)$  and assume that  $\varrho(w)$  is an approximating dictionary for the function  $f(w)$ , so that  $f(w) \approx \varrho(w)' \vartheta$ , with dimension  $M = \dim(\varrho(w))$ . The Lasso estimates for  $\vartheta$  and  $f(w)$  are defined by

$$\hat{\vartheta} \in \arg \min_{t \in \mathbb{R}^M} \sum_{i=1}^n (y_i - \varrho(w_i)' t)^2 + \lambda \sum_{j=1}^M |\hat{\Psi}_j t_j|$$

---

<sup>6</sup>The same is true if only (3.4) is used for variable selection exchanging the roles of (3.3) and (3.4).



$$\hat{f}_{\text{Lasso}}(w) = \varrho(w)' \hat{\vartheta}$$

where  $\lambda$  and  $\hat{\Psi}_j$  are tuning parameters named the penalty level and the penalty loadings. [3] provided estimation methodology as well as results guaranteeing performance for the Lasso estimator under conditions which are common in econometrics including heteroskedastic and non-Gaussian disturbances. Tuning parameters are chosen to balance regularization and bias considerations.<sup>7</sup> Performance bounds for the Lasso, including rates at which  $\sum_{j=1}^L |\vartheta_j - \vartheta|$ ,  $\sum_{j=1}^L |\vartheta_j - \vartheta|^2$ ,  $\sum_{i=1}^n |f(w_i) - \hat{f}(w_i)|^2$  approach zero, are derived on the what is called the Regularization event. The Regularization event is defined by  $\mathcal{R} = \{\lambda > 2c\hat{\Psi}_j^{-1}\mathcal{S}_j \text{ for each } j \leq L\}$ , for a fixed constant  $c > 1$  where  $\mathcal{S}_j$  is the partial derivative of the least squares part of the objective in the  $j$ th direction. Informally, under the regularization event, the penalty level is high enough so that coefficients which cannot be statistically separated from zero are mechanically set to identically zero as a consequence of the absolute values in the above objective function. Because performance bounds are directly proportional to  $\lambda$ , Lasso can be shown to perform well for small values of  $\lambda$  which are nevertheless large enough so that the regularization event occurs with high probability.

Lasso performs particularly well relative to some more traditional regularization schemes (eg. ridge regression) under sparsity: the parameter  $\varrho$  satisfies  $|\{j : \varrho_j \neq 0\}| \leq s$  for some sequence  $s \ll n$ . A feature of the nature of the Lasso penalty that has granted Lasso success is that it sets some components of  $\hat{\vartheta}$  to exactly zero in many cases. Under general conditions,

$$\hat{I} = \{j : \hat{\vartheta}_j \neq 0\} \leq Cs \text{ with probability } 1 - o(1)$$

for a constant  $C$  that depends on the problem. The Post-Lasso estimator is defined as the least squares series estimator that considers only terms selected by Lasso (ie terms with nonzero coefficients):

$$\hat{f}_{\text{Post-Lasso}}(w) = \varrho(w)' \hat{\vartheta}_{\text{Post-Lasso}}; \quad \vartheta_{\text{Post-Lasso}} \in \underset{\{t: t_j=0, \forall j \notin \hat{I}\}}{\operatorname{argmin}} \sum_{i=1}^n (y_i - \varrho(w_i)' t)^2$$

**3.3. Lasso in post-double selection in the additively separable model.** In this section, the use of Lasso is applied directly to the first and second stage problems described in section 3.2 Starting with constructing

---

<sup>7</sup> For the simple heteroskedastic Lasso above, [3] recommend setting

$$\lambda = 2c\sqrt{n}\Phi^{-1}(1 - \gamma/2M), \quad \Psi_j = \sqrt{\sum_{i=1}^n \varrho_j(w_i)^2 (y_i - f(w_i))^2 / n}$$

with  $\gamma \rightarrow 0$  sufficiently slowly, and  $c > 1$ . The choices  $\gamma = \log^{-1} n$  and  $c = 1.1$  are acceptable. The exact values  $f(w_i)$  are unobserved, and so a crude preliminary estimate  $\hat{f}(w_i) = \frac{1}{n} \sum_{i=1}^n y_i$  is used to give  $\hat{\Psi}_j = \sqrt{\sum_{i=1}^n \varrho_j(w_i)^2 (y_i - \hat{f}(w_i))^2}$ . Estimates of  $\hat{f}(w_i)$  can be iterated on as suggested by [3]. The validity of the of the crude preliminary estimate as well as iterative estimates are detailed in the appendix.

an approximation to the operator  $T$ . Each component of  $p$  is regressed onto the dictionary  $q$  giving an approximation for  $Tp_k(x)$  as a linear combination of elements  $q(x)$  for  $1 \leq k \leq K$ . If this can be done with all  $p_k$ , for each  $1 \leq k \leq K$ , then  $T$  applied to a linear combination of  $p(x)$ , namely  $Tp(x)'\beta$ , can also be approximated by a linear combination of elements of  $q$ . The estimation can be summarized with one optimization problem which is equivalent to  $K$  separate Lasso problems. All nonzero components of the solution to the optimization are collected and included as elements of the refined dictionary  $\tilde{p}$ .

$$\hat{\Gamma} = \arg \min_{\Gamma \in \mathbb{R}^{L \times K}} \sum_{k=1}^K \sum_{i=1}^n (p_k(x_i) - q(z_i)'\Gamma_k)^2 + \lambda^{FS} \sum_{k=1}^K \sum_{j=1}^L |\hat{\Psi}_{jk}^{FS} \Gamma_{kj}|.$$

Note that the estimate  $\hat{\Gamma}$  approximates  $T$  in the sense that  $(q(x)'\hat{\Gamma})'\beta$  approximates  $Tp_1(x)'\beta$ . The first stage tuning parameters  $\lambda^{FS}$ ,  $\hat{\Psi}_{jk}^{FS}$  are chosen similarly to the method outlined above but account for the need to estimate effectively  $K$  different regressions. Set

$$\lambda^{FS} = 2c\sqrt{n}\Phi^{-1}(1 - \gamma/2KL),$$

$$\Psi_{jk}^{FS} = \sqrt{\sum_{i=1}^n q_j(z_i)^2 (p_k(x_i) - Tp_k(x_i))^2 / n}.$$

As before, the  $\Psi_{jk}^{FS}$  are not directly observable and so estimates  $\hat{\Psi}_{jk}^{FS}$  are used in their place. The mechanical implementation for calculating  $\hat{\Psi}^{FS}$  is described in the appendix. Details for the constants involved in choosing  $\lambda$  are also given in the appendix. The appearance of  $K$  term in  $\lambda$  ensures that the performance of the Lasso model selection works uniformly well over the  $K$  different Lassos of the first stage.

Running the regression above will yield coefficient estimates of exactly zero for many of the  $\Gamma_{kj}$ . For each  $1 \leq j \leq K$  let  $\hat{I}_k = \{j : \Gamma_{kj} \neq 0\}$ . Then the first stage model selection step selects exactly those terms which belong in the union  $\hat{I}^{FS} = \hat{I}_1 \cup \dots \cup \hat{I}_K$ .

The reduced form selection step proceeds after the first stage model selection step. For this step, let

$$\hat{\pi} = \arg \min_{\pi} \sum_{i=1}^n (y_i - q(x_i)'\pi)^2 + \lambda^{RF} \sum_{j=1}^L |\hat{\Psi}_j^{RF} \pi_j|$$

Where the reduced form tuning parameters  $\lambda^{RF}$ ,  $\hat{\Psi}_{jk}^{RF}$  are chosen according to the method outlined above with

$$\lambda^{RF} = 2c\sqrt{n}\Phi^{-1}(1 - \gamma/2L),$$

$$\Psi_j^{RF} = \sqrt{\sum_{i=1}^n q_j(z_i)^2 (y_i - E[y_i|z_i])^2 / n}.$$

Let  $\hat{I}^{RF} = \{j : \pi_j \neq 0\}$  be the outcome of the reduced form step of model selection.

Considering the set of dictionary terms selected in the first stage and reduced form model selection steps. Let  $\hat{I}$  be the union of all dictionary terms:  $\hat{I} = \hat{I}^{FS} \cup \hat{I}^{RF}$ . Then define the refined dictionary by  $(p(x), \tilde{q}(z)) \equiv (p(x), \{q(z)\}_{j \in \hat{I}})$ . Let  $[P \ \tilde{Q}]$  be the  $n \times (K + |\hat{I}|)$  matrix with the observations of the refined dictionary stacked. The post-double-model selection estimate for  $g(x)$  is defined by

$$\hat{g}(x) = p(x)' \hat{\beta} \quad (3.5)$$

where  $[\hat{\beta}' \ \hat{\eta}']' := ([P \ \tilde{Q}]' [P \ \tilde{Q}])^{-1} [P \ \tilde{Q}]' Y$ .

#### 4. REGULARITY AND APPROXIMATION CONDITIONS

In this section, the model described above is written formally and conditions guaranteeing convergence and asymptotic normality of the Post-Double Selection Series Estimator are given.

**Assumption 1.** *(i)  $(y_i, x_i, z_i)$  are i.i.d. random variables and satisfy  $E[y_i|x_i, z_i] = g(x_i) + h(z_i)$  with  $g \in \mathcal{G}$  and  $h \in \mathcal{H}$  for pre-specified classes of functions  $\mathcal{G}, \mathcal{H}$ .*

The first assumption specifies the model. The observations are required to be identically distributed, which is stronger than the treatment of i.n.i.d variables given in Belloni, Chernozhukov and Hansen (2011).

**4.1. Regularity and approximation conditions concerning the first dictionary.** The following few definitions help characterize smoothness properties of target function  $g$  and approximating functions  $p$ . Let  $f$  be a function defined on the support  $\mathcal{X}$  of  $x$ . Define the Sobolev norm  $|f|_d = \sup_{x \in X} \max_{|a| \leq d} \partial^{|a|} f / \partial x^a$ . In addition, let  $\zeta_d(K) = \max_{|a| \leq d} \sup_{x \in X} \|\partial^{|a|} p(x) / \partial x^a\|$  where  $\|\cdot\|$  denotes the Euclidean norm. Throughout the exposition, all assumptions will be required to hold for each  $n$  with the same set of implied constants.

**Assumption 2.** *There is an integer  $d \geq 0$ , a real number  $\alpha > 0$ , and vectors  $\beta = \beta_K$  such that  $\|\beta\| = O(1)$  and  $|g - p'\beta|_d = O(K^{-\alpha})$  as  $K \rightarrow \infty$ .*

Assumption 2 is standard in nonparametric estimation. It requires that the dictionary  $p$  can approximate  $g$  at a pre-specified rate. Values of  $d$  and  $\alpha$  can be derived for particular classes of functions. [29] gives approximation rates for several leading examples, for instance orthogonal polynomials, regression splines, etc.

**Assumption 3.** *For each  $K$ , the smallest eigenvalue of the matrix*

$$E[(p(x) - Tp(x)(z))(p(x) - Tp(x)(z))']$$

*is bounded uniformly away from zero in  $K$ . In addition, there is a sequence of constants  $\zeta_0(K)$  satisfying  $\sup_{x \in \mathcal{X}} \|p(x)\| \leq \zeta_0(K)$  and  $\zeta_0(K)^2 K/n \rightarrow 0$  as  $n \rightarrow \infty$ .*

This condition is a direct analogue of a combination of Assumption 2 from Newey (1997) and the necessary and sufficient conditions for estimation of partially linear models from [32]. Requiring the eigenvalues of  $E[(p(x) - Tp(x)(z))(p(x) - Tp(x)(z))']$  to be uniformly bounded away from zero is effectively an identifiability condition. It is an analogue of the standard condition that  $E[p(x)p(x)']$  have eigenvalues bounded away from zero specialized to the residuals of  $p(x)$  after conditioning on  $z$ . The second condition of Assumption 3 is a standard regularity condition on the first dictionary.

**4.2. Sparsity Conditions.** The next assumptions concern sparsity properties surrounding the second dictionary  $q(z)$ , used for approximating  $h(z)$ . As outlined above, sparsity will be required along two dimensions in the second dictionary: both with respect to the outcome equation (1) and with respect to the functional  $T$ . Consider a sequence  $s = s_n$  that controls the number of nonzero coefficients in a vector. A vector  $X$  is  $s$ -sparse if  $|\{j : X_j \neq 0\}| \leq s$ . The following give formal restrictions regarding the sparsity of the outcome equation relative to the second approximating dictionary as well as a sparse approximation of the operator  $T$  described above.

**Assumption 4.** *Sparsity Conditions: there is a sequence  $s = s_n$  and  $\phi = s \log(\max\{KL, n\})$  such that*

(i) *Approximate sparsity in the outcome equation: there is a sequence of vectors  $\eta = \eta_L$  that are  $s$ -sparse and the approximation  $\sqrt{\sum_{i=1}^n (h(z_i) - q(z_i)'\eta)^2/n} := \xi_0 = O_P(\sqrt{\phi/n})$  holds. In addition,  $\max_{i \leq n} |h(z_i) - q(z_i)'\eta| = o_P(1)$ .*

(ii) *Approximate sparsity in the first stage. There are  $s$ -sparse  $\Gamma_k = \Gamma_{k,L}$  such that  $\max_{k \leq K} \sqrt{\sum_{i=1}^n (E[p_k(x_i)|z_i] - q(z_i)'\Gamma_k)^2/n} := \xi_{FS} = O_P(\sqrt{\phi/n})$ . In addition,  $\max_{i \leq n, k \leq K} |E[p_k(x_i)|z_i] - q(z_i)'\Gamma_k| = o_P(1)$ .*

(iii)  $s = o(n)$

The assumption above imposes only a mild condition on the sparsity  $s$  and in a sense may be thought of as definitional. In the discussion that follows, additional conditions on the size of the sparsity level  $s$  will be imposed. As a preview, the conditions listed in Assumption 7 will require that  $K\phi n^{-1/2} = Ks \log(\max\{KL, n\})n^{-1/2} \rightarrow 0$ , among other conditions.

The first statement requires that the second dictionary can approximate  $h$  using a small number of terms. The average squared approximation error from using a sparse  $\eta$  must be smaller than the conjectured estimation error

when the subset of the correct terms is known. This restriction on the approximation error follows the convention used by [3]. The second restriction on the maximum approximation error is used to simplify the proofs. The second statement of Assumption 4 generalizes the first approximate sparsity requirement. It requires that each component of the dictionary  $p$  can be approximated by a linear combination of a small set of terms in  $q$ .

Additional discussion of the sparsity assumptions are given in Section 6 which addresses issues arising in the implementation of non-parametric post-double estimates.

**4.3. Regularity conditions concerning the second dictionary.** The following conditions restrict the sample Gram matrix of the second dictionary. A standard condition for nonparametric estimation is that for a dictionary  $P$ , the Gram matrix  $P'P/n$  eventually has eigenvalues bounded away from zero uniformly in  $n$  with high probability. If  $K + L > n$ , then the matrix  $[PQ]'[PQ]/n$  will be rank deficient. However, in the high-dimensional setting, to assure good performance of Lasso, it is sufficient to only control certain moduli of continuity of the empirical Gram matrix. There are multiple formalizations of moduli of continuity that are useful in different settings, see [10], [43] for explicit examples. This paper focuses on a simple condition that seems appropriate for econometric applications. In particular the assumption that only small submatrices of  $Q'Q/n$  have well-behaved eigenvalues will be sufficient for the results that follow. In the sparse setting, it is convenient to define the following sparse eigenvalues of a positive semi-definite matrix  $M$ :

$$\varphi_{\min}(m)(M) := \min_{1 \leq \|\delta\|_0 \leq m} \frac{\delta' M \delta}{\|\delta\|^2}, \quad \varphi_{\max}(m)(M) := \max_{1 \leq \|\delta\|_0 \leq m} \frac{\delta' M \delta}{\|\delta\|^2}$$

In this paper, favorable behavior of sparse eigenvalues is taken as a high level condition and the following is imposed.

**Assumption 5.** *For every constant  $C > 0$  there are constants  $\kappa'' > \kappa' > 0$  which may depend on  $C$  such that with probability  $\rightarrow 1$ , the sparse eigenvalues obey*

$$\kappa' \leq \varphi_{\min}(CsK)(Q'Q/n) \leq \varphi_{\max}(CsK)(Q'Q/n) \leq \kappa''.$$

Assumption 5 requires only that certain “small”  $CsK \times CsK$  submatrices of the large  $p \times p$  empirical Gram matrix  $Q'Q/n$  are well-behaved. This condition seems reasonable and will be sufficient for the results that follow. Informally it states that now small subset of covariates in  $q$  suffer a multicollinearity problem. The could be shown to hold under more primitive conditions by adapting arguments found in [4] which build upon results in [41] and [35]; see also [34].

**4.4. Moment Conditions.** The next conditions are high level conditions about moments and the convergence of certain sample averages which ensure good performance of the Lasso as a model selection device. They allow the use of moderate deviation results given in [22] which ensures good performance of Lasso under non-Gaussian and heteroskedastic errors. [9] discuss plausibility of these types of moment conditions for various models for the case  $K = 1$ . For common approximating dictionaries for a single variable, the condition can be readily checked in a similar manner.

**Assumption 6.** Set  $\epsilon = y - g(x) - h(z)$ . For each  $k \leq K$  let  $W_k = p_k(x) - Tp_k(x)$  and define  $\varpi_{jk}^{FS} = (E[|q_j(z_i)W_{ik}|^3])^{1/3}$ ,  $\varpi_j^{RF} = (E[|q_j(z_i)\epsilon_i|^3])^{1/3}$ . Let  $c, C$  be constants that do not depend on  $n$ . The following conditions are satisfied with probability  $1 - o(1)$  for each  $1 \leq k \leq K, 1 \leq j \leq L$ :

- (i)  $c < E[q_j(z_i)^2 \epsilon_i^2], E[q_j(z_i)^2 W_{ik}^2] < C$
- (ii)  $1 \leq \max_{j \leq L} \Psi_j^{RF} / \min_{j \leq L} \Psi_j^{RF}, \max_{j \leq L} \Psi_{jk}^{FS} / \min_{j \leq L} \Psi_{jk}^{FS} \leq C$
- (iii)  $1 \leq \max_{j \leq L} \varpi_{jk}^{FS} / \sqrt{E\Psi_{jk}^{FS2}}, \max_{j \leq L} \varpi_j^{RF} / \sqrt{E\Psi_j^{RF2}} \leq C$
- (iv)  $\log^3(KL) = o(n)$  and  $s \log(KL) = o(1)$
- (v)  $\max_{k \leq K} \max_{j \leq L} \frac{|\Psi_{jk}^{FS} - \sqrt{E(\Psi_{jk}^{FS})^2}|}{\sqrt{E(\Psi_{jk}^{FS})^2}}, \max_{j \leq L} \frac{|\Psi_j^{RF} - \sqrt{E(\Psi_j^{RF})^2}|}{\sqrt{E(\Psi_j^{RF})^2}} = o(1)$
- (vi)  $\ell\Psi_{jk}^{FS} \leq \widehat{\Psi}_{jk}^{FS} \leq u\Psi_{jk}^{FS}, \ell\Psi_j^{RF} \leq \widehat{\Psi}_j^{RF} \leq u\Psi_j^{RF}$

**4.5. Global Convergence.** The first result is a preliminary result which gives bounds on convergence rates for the estimator  $\widehat{g}$ . They are used in the course of the proof of Theorem 1 below, the main inferential result of this paper. The proposition is a direct analogue of the rates given in Theorem 1 of [29] which considers estimation of a conditional expectation  $g$  without model selection over a conditioning set. The rates obtained in Proposition 1 match the rates in [29].

**Proposition 1.** Under assumptions listed above, the post-double-model-selection estimates for the function  $g$  given in equation 3.5 satisfy

$$\int (g(x) - \widehat{g}(x))^2 dF(x) = O_p(K/n + K^{-2\alpha})$$

$$|\widehat{g} - g|_d = O_P(\zeta_d(n)\sqrt{K}/\sqrt{n} + K^{-\alpha}).$$

## 5. INFERENCE AND ASYMPTOTIC NORMALITY

In this section, formal results concerning inference are stated. Consider estimation of a functional  $a$  on the class of functions  $\mathcal{G}$ . The quantity of

interest,  $\theta = a(g)$ , is estimated by

$$\hat{\theta} = a(\hat{g}).$$

The following assumptions on the functional  $a$  are imposed. They are regularity assumptions that imply that  $a$  attains a certain degree of smoothness. For example, they imply that  $a$  is Fréchet differentiable.

**Assumption 7.** *Either (i)  $a$  is linear over  $\mathcal{G}_1$ ; or (ii) for  $d$  as in Assumption 2,  $\zeta_d(K)^4 K^2/n \rightarrow 0$ . In addition, there is a linear function  $D(f, \tilde{f})$  that is linear in  $f$  and such that for some constants  $C, \nu > 0$  and all  $\tilde{f}, \tilde{f}$  with  $|\tilde{f} - g|_d < \nu$ ,  $|\tilde{f} - g|_d < \nu$ , it holds that  $\|a(f) - a(\tilde{f}) - D(f - \tilde{f}; \tilde{f})\| \leq C(|f - \tilde{f}|_d)^2$  and  $\|D(f; \tilde{f}) - D(f; \tilde{f})\| \leq L|f|_d|\tilde{f} - \tilde{f}|_d$ .*

The function  $D$  is related to the functional derivative of  $a$ . The following assumption imposes further regularity on the continuity of the derivative. For shorthand, let  $D(g) = D(g; g_0)$ .

**Assumption 8.** *Either (i)  $a$  is scalar,  $|D(g)| \leq C|g|_d$ . There is  $\bar{\beta}$  dependent on  $K$  such that for  $\bar{g}(x) = p(x)' \bar{\beta}$ , it holds that  $E[\bar{g}(x)^2] \rightarrow 0$  and  $D(\bar{g}) \geq C > 0$ ; or (ii) There is  $v(x)$  with  $E[v(x)v(x)']$  finite and nonsingular with  $D(g) = E[v(x)g(x)]$  and  $D(p_k) = E[v(x)p_k(x)]$  for every  $k$ . There is  $\tilde{\beta}$  so that  $E[\|v(x) - p(x)' \tilde{\beta}\|^2] \rightarrow 0$ .*

In order to use  $\hat{\theta}$  for inference on  $\theta$ , an approximate expression for the variance  $\text{var}(\hat{\theta})$  is necessary. As is standard, the expression for the variance will be approximated using the delta method. An approximate expression for the variance of the estimator  $\hat{\theta}$  therefore requires an appropriate derivative of the function  $a$ , (rather, an estimate). Let  $A$  denote the derivatives of the functions belonging to the approximating dictionary,  $A = (D(p_1), \dots, D(p_K))'$ . Let  $\hat{A} = \frac{\partial a(p(x)' \hat{b})}{\partial b}(\hat{\beta})$ . The approximate variance, from the delta method is given by  $V = V_K$ :

$$\begin{aligned} V &= A Q^{-1} \Sigma Q^{-1} A \\ \Omega &= E[(p(x) - T p(x))(p(x) - T p(x))'] \\ \Sigma &= E[(p(x) - T p(x))(p(x) - T p(x))'(y - g(x))^2] \end{aligned}$$

These quantities are unobserved but can be estimated:

$$\begin{aligned} \hat{V} &= \hat{A} \hat{\Omega}^{-1} \hat{\Sigma} \hat{\Omega}^{-1} \hat{A} \\ \hat{\Omega} &= \sum_{i=1}^n (p(x_i) - \hat{p}(x_i))(p(x_i) - \hat{p}(x_i))' / n \\ \hat{\Sigma} &= \sum_{i=1}^n (p(x_i) - \hat{p}(x_i))(p(x_i) - \hat{p}(x_i))'(y - \hat{g}(x_i))^2 / n \end{aligned}$$



The elements  $\widehat{p}_k(x_i)$  are obtained as the predictions from the least squares regression of  $p_k(x_i)$  onto the selected  $\widehat{q}(z_i)$ . Then  $\widehat{V}$  is used as an estimator of the asymptotic variance of  $\widehat{\theta}$  and assumes a sandwich form.

The following assumptions are needed in order to bound  $\widehat{V} - V$ :

**Assumption 9.** Define the moments  $\varpi_j^{q\epsilon} = (E[|q_j(z_i)\epsilon_i|^3])^{1/3}$ ,  $\varpi_k^{W\epsilon} = (E[|W_{ki}\epsilon_i|^3])^{1/3}$ . Let  $q > 4, C > 0$  be constants. Then for  $k \leq K, j \leq L$

- (i)  $E[|\epsilon_i|^q | x_i, z_i] \leq C$
- (ii)  $\varpi_{jk}^{q\epsilon} / \sqrt{E[q_j(z_i)^2 \epsilon_i^2]}, \varpi_k^{W\epsilon} / \sqrt{E[W_{ki}^2 \epsilon_i^2]} \leq C$
- (iii)  $n^{2/q} K \phi / \sqrt{n} = o(1)$
- (iv)  $\left( \zeta_0(K) \sqrt{K} / \sqrt{n} \right) \left( s \sqrt{K/n} + \sqrt{n} K^{-\alpha} \right) \max_{i,j} |q_j(z_i)| = o_P(1)$

The assumption that  $q > 4$  moments of  $\epsilon$  are bounded slightly strengthens the condition in [29] that fourth moments are bounded. The condition is necessary for the estimation of the final standard errors. Similarly, condition (iii) is stronger than the rate condition listed in Assumption 3. The more stringent condition is also useful for estimating standard errors. Finally, condition (ii) is analogous to Assumption 7, condition (iii) and is again useful for controlling the tail behavior of certain self-normalized sums.

The next result is the main result of the paper. It establishes the validity of standard inference procedure after model selection as well as validity of the plug in variance estimator.

**Theorem 1.** Under the Assumptions 1-7,9 and Assumption 8(i), and in addition  $\sqrt{n} K^{-\alpha} \rightarrow 0$  then  $\widehat{\theta} = \theta + O_P(\zeta_d(K) \sqrt{n})$  and

$$\sqrt{n} V^{-1/2} (\widehat{\theta} - \theta) \xrightarrow{d} N(0, 1), \quad \sqrt{n} \widehat{V}^{-1/2} (\widehat{\theta} - \theta) \xrightarrow{d} N(0, 1)$$

If Assumptions 1-7,9 and Assumption 8(ii) hold with  $d = 0$  and in addition  $\sqrt{n} K^{-\alpha} \rightarrow 0$  then for  $\bar{V} = E[v(x)v(x)' \text{var}(y|x)]$ , the following convergences hold.

$$\sqrt{n} (\widehat{\theta} - \theta) \xrightarrow{d} N(0, \bar{V}), \quad \|\widehat{V} - \bar{V}\| \xrightarrow{p} 0$$

The theorem shows that the outlined procedure gives a valid method for performing inference for functionals after selection of series terms. Note that under assumption 8(i) the  $\sqrt{n}$  rate is not achieved because the functional  $a$  does not have a mean square continuous derivative. By contrast, Assumption 8(ii) is sufficient for  $\sqrt{n}$ -consistency. Conditions under which the particular assumptions regarding the approximation of  $g$  hold are well known. For example, conditions on  $K$  for various common approximating dictionaries including power series or regression splines etc follow those directly derived in [29]. Asymptotic normality of these types of estimates under the high

dimensional additively separable setting should therefore be viewed as a corollary to the above result.

Consider one example with the functional of interest being evaluation of  $g$  at a point  $x^0$ :  $a(g) = g(x_0)$ . In this case,  $a$  is linear and  $D(\bar{g}) = \bar{g}(x_0)$  for all functions  $\bar{g}$ . This particular example does not attain a  $\sqrt{n}$  convergence rate provided there is a sequence of functions  $g_K$  in the linear span of  $p = p^K$  such that  $E[g_K(x)^2]$  converges to zero but  $g_K(x_0)$  is positive for each  $K$ . Another example is the weighted average derivative  $a(g) = \int w(x) \partial g(x) / \partial x$  for a weight function  $w$  which satisfies regularity conditions. For example, the theorem holds if  $w$  is differentiable, vanishes outside a compact set, and the density of  $x$  is bounded away from zero wherever  $w$  is positive. In this case,  $a(g) = E[v(x)g(x)]$  for  $v(x) = -f(x)^{-1} \partial w(x) / \partial x$  by a change of variables provided that  $x$  is continuously distributed with non vanishing density  $f$ . These are one possible set of sufficient conditions under which the weighted average derivative does achieve  $\sqrt{n}$ -consistency.

## 6. ADDITIONAL DISCUSSION OF IMPLEMENTATION

The theorem above states that for a fixed (nonrandom) sequence  $K = K_n$ , asymptotically normal estimates are achieved for certain functionals of interest of  $g(x)$  under the right regularity conditions. In practice, the choice of  $K_n$  is important and it is useful to have a data-driven means by which to choose  $K_n = \hat{K} = \hat{K}(\{(x_i, y_i)\}_{i=1}^n)$ . This section provides suggestions for choosing such  $\hat{K}$ . This paper leaves these suggestions as heuristics and does not derive formal theory for their asymptotic performance; however, the finite sample performance of these heuristics is explored in the simulations in Section 7.<sup>8</sup>

Suppose that candidates for  $K$  belong to the integer set  $\{\underline{K}, \dots, \overline{K}\}$ . Suppose that  $L = L_n$  is nonrandom as before and does not vary with  $K \in \{\underline{K}, \dots, \overline{K}\}$ . A simple proposal is as follows: for each  $K$ , construct using the post-double model selection routine, a reduced second dictionary  $\tilde{q}_K^L$ . This results in a set of selected dictionaries which are candidates for a final estimation step:

$$\left\{ (p^K(x), \tilde{q}_K^L(z)), \dots, (p^{\overline{K}}(x), \tilde{q}_{\overline{K}}^L(z)) \right\}.$$

Then  $\hat{K}$  can be chosen by selecting from the dictionaries in the above set. In principal this can be done in many ways, including cross validation or BIC, possibly with additional over-smoothing (i.e. choosing  $\hat{K}$  larger than say the mean-square error optimal.)

---

<sup>8</sup>There is also the question of whether the penalty levels  $\lambda^{FS}$  and  $\lambda^{RF}$  in the Lasso optimizations can be chosen in a more data-driven way. There is less flexibility in these choices, since the Lasso bounds are predicated on the Regularization event. Using a smaller penalty level than suggested leads to over-selection of control variables, which can bias estimates. Further discussion of the effects of post-model-selection inference with over-selection can be found in [7].

The informal reasoning behind this proposal is that for each  $K$ , the post-double model selection selects confounders in a way so that the omitted variables bias resulting from possibly omitting covariates predictive of elements of  $p^K(x)$  is small relative to sampling variability. But there can be another component of omitted variables bias: namely, the omitted variables biases resulting in excluding all confounders predictive only of signal in  $x$  which is not accounted by  $p(x)$ . However, this second component of omitted variables bias will plausibly be small if  $K$  is chosen appropriately, (i.e. whenever  $\sqrt{n}K^{-\alpha} \rightarrow 0$ , a similar bound on such omitted variables bias may be expected to hold.)

In addition to issues arising from choosing  $K$ , the sparsity conditions imposed in the previous sections are restrictive. However, without such a sparse structure, it is difficult to construct meaningful estimates of  $g(x)$ . Furthermore, at the current time, there are no widely used procedures to test the null hypothesis of a sparse model that the author is aware of.

A major restriction implicit in the sparsity assumption is that the sparse approximation errors are small for *all* terms in  $p(x)$ . For instance, if  $p(x) = (x, x^2, x^3, \dots)$ , it is much less demanding to ask that there exists a good sparse predictor of  $x$  based on  $q(z)$ , than it is to ask that  $x, x^2, x^3 \dots$  *all* have quality sparse predictors. On the other hand, it is possible that the transformations  $x^2 + x^3$  or  $x^3 - x^2$  might have much better sparse representations in terms of  $q(z)$  than  $x^2$  and  $x^3$  have individually.

Therefore, an alternative strategy for the first stage is to have a model selection step for many distinct linear combinations  $\alpha'p(x)$  given by  $\alpha \in \mathcal{A}$ . The strategy is outlined as follows. First, gather the set  $\{\alpha'p(x) : \alpha \in \mathcal{A} \subset \mathbb{R}^K\}$  into an extended first stage dictionary:  $p_{FS}(x)$ . Select a reduced conditioning dictionary  $\tilde{q}(z)$  with the nonparametric post-double selection method described above, except using  $p_{FS}(x)$  in the first stage model selection step. Finally, in the post-model-selection estimation step, estimate using  $(p(x), \tilde{q}(z))$ . This strategy is potentially useful since it further reduces the possibility for omitted variables bias. A clear tradeoff with using a distinct first stage dictionary is that due to the additional model selection steps introduced, more variables from  $p(z)$  can potentially be selected, leading to higher variability of the final estimate of  $g(x)$ .

As with the data-driven choice of  $K$ , this suggestion is kept at the level of a heuristic at this moment. However, arguments in the proofs of the main results can easily be extended to allow an extended to allow a first stage dictionary  $p_{FS}(x)$  provided that it has a subdictionary,  $p^K(x)$  for which Assumptions 1-4 hold, and that the number of selected conditioning variables  $\tilde{q}(z)$  remains  $O(sK)$  with high probability. For example, if  $\dim(p_{FS}(x)) \leq C\dim(p(x))$  then no substantive modification to the proof are necessary.

The finite sample performance of these heuristics is explored in the simulations in Section 7.

## 7. SIMULATION STUDY

The results stated in the previous section suggest that post double selection type series estimation should exhibit good inference properties for additively separable conditional expectation models when the sample size  $n$  is large. The following simulation study is conducted in order to illustrate the implementation and performance of the outlined procedure. Results from several other candidate estimators are also calculated to provide a comparison between the post-double method and other methods. Estimation and inference for two functionals of the conditional expectation function are considered. Two simulation designs are considered. In one design, the high dimensional component over which model selection is performed is a large series expansion in four variables. In the other design, the high dimensional component is a linear function of a large number of different covariates.

**7.1. Low Dimensional Additively Separable Design.** Consider the following model of continuous variables  $x, z$  of form:

$$E[y|x] = E[y|x, z] = g(x) + h(z)$$

where in this simulation, the true function of interest,  $g(x)$ , and the conditioning function  $h(z)$  are given by :

$$g(x) = \text{logistic}(x) - \frac{1}{2}$$

$$h(z) = \text{logistic}\left(\sum_{j=1}^{\dim(z)} z_j\right) - \frac{1}{2}$$

where  $\text{logistic}(x) = \frac{\exp(x)}{1+\exp(x)}$  and the  $\frac{1}{2}$  terms in the expression for  $g$  is used to ensure identifiability via  $g(0) = 0$ . Ex post, the function is simple, however, for the sake of the simulation, knowledge of the logistic form is assumed unknown. Importantly, the logistic function will not belong exactly in the span of any finite series expansion used in the below simulation. The second function  $h$  is similar, being defined by a combination of a logistic function of a linear combination of the  $z$  variables. The logistic part can potentially require many interaction terms unknown in advance to produce an accurate model. The component functions  $g$  and  $h$  will be used throughout the simulation. The remaining parameters, eg. dictating the data generating processes for  $(y, x, z)$  will be changed across simulation to give an illustration of performance across different settings.

The objective is to estimate a population average derivative,  $\theta^{(1)}$ , and a function evaluation,  $\theta^{(2)}$  given by

- (i)  $\theta_1 = a_1(g) = \int_{\text{supp}(x)} \frac{\partial g}{\partial x}(x) dF(x)$
- (ii)  $\theta_2 = a_2(g) = g(\text{quantile}(x, .75)) - g(\text{quantile}(x, .25)).$

$\hat{\theta}_1$  and  $\hat{\theta}_2$  and estimates of standard errors are obtained using the post-double methodology outlined in the paper with approximating dictionaries specified below. The plug in estimate of the average derivative,  $\theta_1$ , is integrated against the empirical distribution of  $x$ .

The covariates and outcome are drawn as follows. The distribution of the conditioning set  $z$  is set at  $z \sim N(0, S_{\dim(z)}^{1/2})$  where  $S_{\dim(z)}^{1/2}$  is the Toeplitz matrix of size  $\dim(z)$  with decay base of  $\frac{1}{2}$ ; therefore, the correlations between  $z_j$  and  $z_k$  are set to  $\text{corr}(z_j, z_k) = (\frac{1}{2})^{-|j-k|}$ . The variable of interest,  $x$  is determined by  $x = h(z) + v$  with  $v \sim N(0, 1) \cdot \sigma_v$  independent of  $z$ . The structural errors  $\epsilon = y - E[y|x]$  are drawn  $N(0, 1) \cdot \sigma_\epsilon$  independent of  $x$ . Several simulations of this model are conducted by varying the sample size  $n$ , the dependence between  $x$  and the remaining regressors  $z$ , as well as the size of the residual errors  $\epsilon$ . The sample size varied and is set to  $n \in \{500, 1000\}$ . The dependence between  $x$  and the remaining covariates is dictated by  $\sigma_v$ . To capture high and low dependence between the covariates, the values  $\sigma_v \in \{1, 2\}$  are used. Finally, the variability of the structural shocks are set to  $\sigma_\epsilon \in \{1, 2\}$ .

Estimation is based on series expansion using Hermite polynomials.  $K$  is set to  $K = \text{floor}(n^{1/3})$ . Therefore,  $g(x)$  is approximated using a  $K$  order polynomial. The series expansion  $q(z)$  for the function  $h(z)$  spans all polynomials in  $z$  of order  $\leq K$ . Specifically, using interactions (products) of univariate  $m$ -order Hermite polynomials,  $H_m$ , on single components of  $z$ , the elements of  $q(z)$  consist of terms in  $\left\{ \prod_{j=1}^{\dim(z)} H_{m_j}(z_j) : \sum_{j=1}^{\dim(z)} m_j \leq K \right\}$ .

In addition to the standard post-double lasso based model, several alternative estimates are calculated for comparison. We give three ad hoc estimators which seem to be sensible data-driven ways to choose the number of series terms (see the description and discussion of these in Section 6). The first is the Post-Double selection where  $K = \hat{K}$  is chosen from a set of values  $\text{floor}(\frac{1}{2}n^{1/3}) \leq \hat{K} \leq \text{floor}(2n^{1/3})$ . The choice is by the following procedure. First, run post-double selection for each choice of  $K$ . This produces a set of candidate models  $\{(p^K, \tilde{q}_K^L), \dots, (p^{\bar{K}}, \tilde{q}_{\bar{K}}^L)\}$  which can be compared each other. In this simulation, the preferred method of comparison is BIC since it is simple computationally, relative, for example, to cross-validation. Then the final value  $\hat{K}$  is chosen to be  $\hat{K} = K_{\text{BIC}} + 1$ . This estimator is referred to as the Post-Double Set estimator in the simulation results tables.

In addition, the second suggestion from Section 6, which is to augment the first dictionary in the first stage model selection step, is considered. For given  $p(x)$ , the extended first stage dictionary  $p_{FS}$  is constructed by

$$p_{FS}(x) = p(x) \cup \{p_j(x) + p_{j'}(x) : j < j'\} \cup \{p_j(x) - p_{j'}(x) : j < j'\}$$

which is used in place of  $p(x)$  in the first stage model selection step. This gives for fixed  $K$  an estimate which is called the Post-Double Ext estimate in the simulation results tables.

Finally, we consider a hybrid of the Post-Double Set estimator and the Post-Double Ext estimator where a model is selected using the Post-Double-Extend method for each  $K \in \{\underline{K}, \dots, \overline{K}\}$ . These models are compared and the choice  $\hat{K}$  is made according to  $\hat{K} = K_{\text{BIC}} + 1$ . This estimator is called the Post-Double Set+Ext estimator in the simulation results tables.

In addition to Post-Double-based estimators, two additional series estimators which are designed to approximate the function  $g(x) + h(z)$  are compared. The first is based on a series approximation which assumes that  $h(z)$  is additively separable so that  $h(z) = \sum_{j=1}^{\dim(z)} h_j(z_j)$ . The approximating series for each term  $h_j(z_j)$  consist of Hermite polynomials in  $z_j$  of order  $\leq K$ . The second series estimator uses identical series as the post-double selection model, but performs no model selection; that is, it proceeds under the same mechanical procedure as would be used for a standard series estimator and uses the Moore-Penrose pseudo-inverse if it needs to invert a singular matrix. Second, a single step selection estimator is provided. The single step estimator is done by performing a first stage lasso on the union of the two dictionaries, then re-estimating coefficients of the remaining dictionary terms in the second stage. Finally, an infeasible estimator is provided, where estimation proceeds as standard series estimation given the dictionary  $p$  and as if  $h(x_i)$  were known for each  $i$ .

Results for estimating  $\theta_1$  and  $\theta_2$  are based on 500 simulations for each setting described earlier. For each estimator, the median bias, median absolute deviation, and rejection probability for a 5-percent level test of  $H_{01} : \theta_1 = \theta_{01}$  or  $H_{02} : \theta_2 = \theta_{02}$  are presented. Results for estimating  $\theta_1$  are presented in Table 5.4. In each design, estimates of  $\theta$  based on post double selection exhibit small median absolute deviation relative to the competing estimates. With the exception of the infeasible estimates the post double selection estimates are also the only estimates which exhibit reasonable rejection frequencies consistently across all settings. Results for estimation of  $\theta_2$  are reported in Table 5.5. The results are qualitatively similar to those for  $\theta_1$ . The only reasonable rejection frequencies are obtained with the post-double selection. A small amount of size distortion can be seen as rejection frequencies are closer to 10-percent in most simulations. The distortion in the post double estimator matches that in the infeasible estimator suggesting that they are driven by bias in approximating a nonlinear function with a small bias rather than a consequence of model selection.

**7.2. High Dimensional Additively Separable Design.** In this design, a high dimensional setting is considered:

$$E[y|x, z] = E[y|x, z] = g(x) + h(z).$$

For the functions

$$g(x) = \text{logistic}(x) - \frac{1}{2}$$

$$h(z) = \sum_{j=1}^{\dim(z)} \left(\frac{1}{2}\right)^{j-1} z_j$$

In this model, the dimension of  $z$  is large compared to the sample size and is set to  $\dim(z) = \text{floor}(2n)$  for each scenario. The variables  $z$  are drawn as above from a Gaussian with  $z \sim N\left(0, S_{\dim(z)}^{1/2}\right)$  so that, again, the correlation structure is  $\text{corr}(x_j, x_k) = \left(\frac{1}{2}\right)^{-|j-k|}$ . The target function  $g(x)$  remains the same as before. The dependence between  $x$  and  $z$  is defined with  $x = h(z) + v$ . The specifications for  $v$ ,  $\sigma_v$ ,  $\epsilon$ , and  $\sigma_\epsilon$  are the same as they were in the low dimensional example. Estimation is again based on a series expansion in terms of Hermite polynomials in  $x$ .  $K$  is taken to be  $K = \text{floor}(n^{1/3})$ . The second dictionary comprises of the variables  $z_j$  themselves. The simulation evaluates the performance of the estimator when the goal is to control for many distinct possible sources of confounding. Results are recorded in Table 5.6 for sample size  $n = 500$  and Table 5.7 for  $n = 1000$ . In this simulation, the single selection and infeasible estimators are defined as they were in the low dimensional simulation. The series estimator (Series I) uses the first  $\text{floor}(4n/5)$  covariates set of controls in estimation. The second series estimator (Series II) randomly selects  $\text{floor}(4n/5)$  of the covariates to use in estimation. The results are qualitatively similar to those for the first design. Relative to other methods, the post-double selection method exhibits substantially lower bias. The rejection frequencies obtained with the post-double selection are considerably closer to the target 5% level than with other methods.

## 8. EMPIRICAL APPLICATION: ESTIMATING THE EFFECT OF EXPORT ON HIV

The results in the preceding sections show how variable selection methods can be used to estimate additively separable models, in which the component of interest can be considered effectively randomly assigned conditional on observables. This section illustrates use of the results by reexamining [30], which studies of the impact of trade on HIV incidence rates in a sample of African countries. The original results are briefly reviewed, after which estimates using the methods developed in this paper are calculated and discussed.

The arguments presented in [30] propose two causal mechanisms relating trade to HIV incidence rates in African countries. The first is simply that increased income provided by more exports leads to an increase in risky behavior. The second mechanism is that increased trade leads to an increase in trucking and movement of people who engage in risky behavior. The basic



problem in estimating the causal impact of trade on HIV is that country-level export rates are not randomly assigned. It is likely that there are factors like government specific policies that are associated to both trade rates and HIV rates.

The baseline model estimated in [30] is for country-level incidence rates running from 1985 to 2007, which conditions on country level fixed effects, time level effects as well as lagged HIV prevalence.<sup>9</sup> In the present analysis, the argument given in [30], that the export rates defined above may be taken as exogenous relative to HIV incidence once observables and government or policy variables are controlled for, is taken for granted. While [30] controls for these possible confounds by including country-specific and year-specific fixed effects and lagged prevalence, this paper allows a much richer set of variables to be used as controls. All of the new control variables are described below and can be conveniently constructed from the original dataset used by [30]. An additional benefit is that selection of a conditioning set offers a very attractive complimentary analysis to the standard practice of robustness checks. Robustness checks typically estimate several perturbations of the baseline model. If the value of causal estimates are insensitive to these perturbations, then this is considered evidence in favor of the researcher's original conclusions. While robustness checks are useful, it is very difficult to come up with a principled means for choosing a correct set of perturbed models to use. The methods here give one such formalized procedure and serve to compliment the robustness checks performed in [30].

This paper considers a model specified by:

$$y_{it} = g(x_{it}) + \alpha_i + \gamma_t + w'_{it}\delta + z'_{it}\theta + \varepsilon_{it}$$

where  $i$  indexes country,  $t$  indexes times,  $\alpha_i$  are country-specific effects that control for any time-invariant country-specific characteristics,  $\gamma_t$  are time-specific effects that control flexibly for any aggregate trends,  $w_{it}$  is the lagged HIV prevalence rate. The outcome  $y_{it}$  is log measure of HIV incidence calculated in two different ways. The first measure, named UNAIDS-based incidence, is based on data from UNAIDS, the United Nations organization responsible for reporting on the global HIV epidemic. The second way, named Death-based incidence, extrapolates backwards from mortality rates from the Demographic Yearbook Historical Supplement. The  $x_{it}$  are logs of three different export measures: total export value reported by the World Development Indicators; total export value reported by NBER - United Nations Trade Data; and total export volume reported by NBER-United Nations Trade Data. The measures are labeled Log Value (WDI), Log Value

---

<sup>9</sup>The arguments in this paper, in conjunction with those given for selection of controls in high dimensional panel models in [7], could be used to justify allowing dependence within countries over time. The standard errors in the current analysis are clustered by country whereas [30] assumes an AR(1) structure and a linear model and proceeds with Prais-Winston regression.

(NBER) Value, and Log Volume (NBER). For details, please refer to [30]. This paper abstracts away from important issues arising from measurement error in both incidence rates and export rates in order to provide a clearer illustration of the model selection methods.

The set of controls represented by  $z_{it}$  is richer than the controls considered in the original paper.  $z_{it}$  is constructed from interactions of an aggregate linear time trend,  $t$ , an aggregate quadratic time trend,  $t^2$ , aggregate periodic time trends with periods 4 and 8 years,  $\sin(2\pi t/4)$ ,  $\cos(2\pi t/4)$ ,  $\sin(2\pi t/8)$ ,  $\cos(2\pi t/8)$ , log-population, indicators for larger regions in Africa, initial values of HIV prevalence, GDP, log-population, and export levels. The final set of controls consists of 78 variables and can be obtained by request. The samples consists of between 720-747 observations for the UNAIDS-based incidence measure and 161-166 for the Death-based incidence measure, with discrepancies arising from missing observations. Therefore, a conditioning set of 78 variables, though more robust than the baseline, is large enough to potentially cause statistical problems. Therefore, this analysis will apply the model selection procedure discussed in the earlier sections. The estimates of  $g(x_{it})$  are based on a series expansion for  $g$  in terms of Hermite polynomials. The order of the approximating polynomial is chosen by  $K = \text{floor}(n^{1/4})$ , where  $n$  is the total number of observations (not observational units).

Estimates for the sample average derivative,  $\frac{1}{n} \sum \hat{g}'(x_{it})$ , based on the UNAIDS measure are presented in Table 5 and estimates based on Death-based incidence are presented in Table 6. In each table, the first panel uses the Log Value (WDI) measure, the second panel uses the Log Value (NBER) measure, and the third panel uses the Log Volume (NBER) measure. The estimate is interpreted as the average percent increase in HIV incidence resulting from an exogenously given percent increase in export over the sample export values. The tables present estimates and 95% confidence intervals, as well as lists of selected variables for each analysis. In addition to the post-model selection estimates, a baseline estimate, corresponding to the baseline model in [30] (excluding  $z_{it}$ , yielding  $y_{it} = g(x_{it}) + \alpha_i + \gamma_t + w'_{it}\delta + \varepsilon_{it}$ ), and a model using the full set of controls without model selection are presented.

The estimates presented in Table 5 are not significantly different from 0 in any of the circumstances. This is partly consistent with the findings in [30], in that the significance does not withstand the robustness checks. The estimates here are, in principal, more variable relative to [30] since this paper considers non-parametric specifications for  $g$  whereas [30] considers a linear specification. The estimates in Table 6 are considerably more interesting. In particular, the interpretation of the estimates changes depending on which set of controls is used. Using the first measure of export, Log Value (WDI), the estimated average derivative is significantly different from zero under the baseline model. However, the estimate fails to maintain significant separation from zero when the full set of controls is used. This can be because the

true causal effect based on the richer conditioning set is null, or because the inclusion of many control variables presents a serious limitation in terms of statistical precision. In this situation, using a selected set of controls yields a statistically significant, positive estimate. In addition, the corresponding confidence interval is more narrow. Therefore, for these particular measures of incidence and export, the model selection method supports the conclusion of the baseline estimate of a positive average effect. Using the other two measures of export, the interpretations of the estimates, presented in Panels 2-3, do not change between the three estimation strategies.

## 9. CONCLUSION

This paper considers the problem of choosing a parsimonious conditioning set in the context of nonparametric regression. Convergence rates and inference results are provided for series estimators of additively separable models with a high dimensional component. Lasso continues to have good selection properties in this context and can be used in post-model selection inference when two model selection steps are performed.

## APPENDIX A. LASSO PENALTY LOADINGS IMPLEMENTATION

For completeness, this section presents implementation details for Cluster-Lasso. The details are the same as those given in [3]. Following a mechanical description of penalty loadings choice, conditions are presented which are sufficient for establishing the asymptotic validity of the proposed algorithm in this appendix.

Feasible options for setting the penalty level and the loadings for  $j = 1, \dots, L$  are

$$\begin{aligned}
 \text{Initial:} \quad & \hat{\Psi}_{jk}^{FS} = \sqrt{\frac{1}{n} \sum_{i=1}^n q_j(z_i)^2 [p_k(x_i) - \frac{1}{n} \sum_{i=1}^n p_k(x_i)]^2} \\
 & \hat{\Psi}_j^{RF} = \sqrt{\frac{1}{n} \sum_{i=1}^n q_j(z_i)^2 [y_i - \frac{1}{n} \sum_{i=1}^n y_i]^2} \\
 \\ 
 \text{Refined:} \quad & \hat{\Psi}_{jk}^{FS} = \sqrt{\frac{1}{n} \sum_{i=1}^n q_j(z_i)^2 \widehat{W}_{ik}^2}, \\
 & \hat{\Psi}_j^{RF} = \sqrt{\frac{1}{n} \sum_{i=1}^n q_j(z_i)^2 \widehat{\varepsilon}_i^2} \\
 \\ 
 \text{Penalty:} \quad & \lambda^{FS} = 2c\sqrt{n}\Phi^{-1}(1 - \gamma/(2KL)) \\
 & \lambda^{RF} = 2c\sqrt{n}\Phi^{-1}(1 - \gamma/(2L))
 \end{aligned}$$

where  $c > 1$  is a constant,  $\gamma \in (0, 1)$ ,  $\Phi$  is the cumulative distribution function for the standard Gaussian distribution, and  $\widehat{\varepsilon}_i$  is an estimate of  $\varepsilon_i := y_i - E[y_i|z_i]$ . Let  $N_{loadings} \geq 1$  denote a bounded number of iterations. This paper uses  $c = 1.1$ ,  $\gamma = 0.1/\log(\max\{KL, n\})$ , and  $N_{loadings} = 15$  in simulation examples. In what follows, Lasso/Post-Lasso estimator indicates that the practitioner can apply either the Lasso or Post-Lasso estimator. The simulations and empirical example in this paper use Post-Lasso.

### Algorithm of Cluster-Lasso penalty loadings

- (1) Specify penalty loadings according to the initial option above. Use these penalty loadings in computing the Lasso/Post-Lasso estimators defined in Section 3.3. Then compute residuals  $\widehat{W}_{ik} = p_k(x_i) - q(z_i)' \widehat{\Gamma}_k$ ,  $\widehat{\epsilon}_i = y_i - q(z_i)' \widehat{\theta}$  for  $i = 1, \dots, n$  and  $k = 1, \dots, K$ .
- (2) If  $N_{penalty} > 1$ , update the penalty loadings according to the refined option above and update the Lasso/Post-Lasso estimator. Then compute a new set of residuals using the updated Lasso/Post-Lasso coefficients  $\widehat{W}_{ik} = p_k(x_i) - q(z_i)' \widehat{\Gamma}_k$ ,  $\widehat{\epsilon}_i = y_i - q(z_i)' \widehat{\theta}$  for  $i = 1, \dots, n$  and  $k = 1, \dots, K$ .
- (3) If  $N_{loadings} > 2$ , repeat step (2)  $N_{loadings} - 2$  times.  $\square$

The results of Proposition 1 and Theorem 1 in this paper rely on asymptotic validity of penalty loadings in the sense that  $\ell \Psi_j^{RF} \leq \widehat{\Psi}_j^{RF} \leq u \Psi_j^{RF}$  for every  $j$  and  $\ell \Psi_{jk}^{FS} \leq \widehat{\Psi}_{jk}^{FS} \leq u \Psi_{jk}^{FS}$  for every  $j, k$  with probability  $1 - o(1)$ ,  $\ell \xrightarrow{P} 1$ , and  $u \leq C < \infty$ . [3] list several primitive assumptions which imply asymptotic validity when  $K = 1$ . The modifications required on their conditions are not substantive and therefore, this paper assumes asymptotic validity of the penalty loadings as a high level condition.

## APPENDIX B. PROOFS OF THE MAIN RESULTS

**B.1. Additional notation used in proofs.** In the course of the proofs, the following notation will be used. First, we use the standard stacking convention for random variables indexed by  $i = 1, \dots, n$  into a column vector of size  $n \times 1$  so that  $y_i$  are stacked in  $Y$ ,  $g_i = g(x_i)$  are stacked in  $G$ ,  $h_i = h(z_i)$  are stacked in  $H$ , and so forth. Let  $\widehat{I}$  be the full set of series terms chosen in the final estimation coming from the dictionary  $q$ .  $\widehat{I}$  is given by  $\widehat{I} = \widehat{I}_0 \cup \widehat{I}_{R,F} \cup \widehat{I}_1 \cup \dots \cup \widehat{I}_K$ . Define for any subset  $J \subset [p]$ ,  $Q[J]$  to be the corresponding set of selected dictionary elements. Let  $b$  be the least squares coefficient for the regression of any vector  $U$  on  $Q[J]$  so that  $b = b(U; J) = (Q[J]'Q[J])^{-1}Q[J]'U$ . Let  $\mathcal{P}_{\widehat{I}} = Q[\widehat{I}](Q[\widehat{I}]'Q[\widehat{I}])^{-1}Q[\widehat{I}]$  be the sample projection onto the space spanned by  $Q[\widehat{I}]$ . Let  $\mathcal{M}_{\widehat{I}} = I_n - \mathcal{P}_{\widehat{I}}$  be projection onto the corresponding orthogonal subspace. Let  $\widehat{\Omega} = P'\mathcal{M}_{\widehat{I}}P/n$ . Let  $\Omega = E[(p(x) - Tp(z))(p(x) - Tp(z))']$ . Decompose  $P = m + W$  where the  $i$ th element of  $m$  is defined by  $m_i = E[p(x_i)|z_i]$ . Let  $\bar{\Omega} = W'W/n$ . Let  $\|\cdot\|$  denote Euclidean norm when applied to a vector and the matrix norm  $\|A\| = \sqrt{\text{tr}A'A}$  when applied to a square matrix. Let  $\|\cdot\|_1$  and  $\|\cdot\|_\infty$  denote  $L_1$  and  $L_\infty$  norms. Let  $\xi_{FS} = \max_k \sqrt{(m_k - Q\Gamma_k)'(m_k - Q\Gamma_k)/n}$  and  $\xi_{RF} = \sqrt{(G + H - Q\pi)'(G + H - Q\pi)/n}$  be the approximation error in the first stage and reduced form.

**B.2. Proof of Proposition.** Begin by establishing the claim  $\|\hat{\Omega} - \Omega\| \xrightarrow{P} 0$  by bounding each of the following terms separately:  $\|\hat{\Omega} - \Omega\| = \|\hat{\Omega} - \bar{\Omega} + \bar{\Omega} - \Omega\| \leq \|\hat{\Omega} - \bar{\Omega}\| + \|\bar{\Omega} - \Omega\|$ . The argument in Theorem 1 of Newey (1997), along with the fact that  $\sup_{x \in \mathcal{X}} \|p(x) - Tp(x)\| \leq 2 \sup_{x \in \mathcal{X}} \|p(x)\|$  gives the bound  $\|\bar{\Omega} - \Omega\| = O_P(\zeta_0(K)K^{1/2}/\sqrt{n})$ . Next bound  $\|\hat{\Omega} - \bar{\Omega}\|$ . Using the decomposition,  $P = m + W$ , write  $\hat{\Omega} = (m + W)' \mathcal{M}_{\hat{I}}(m + W)/n = W'W/n - W'(I_n - \mathcal{M}_{\hat{I}})W/n + m' \mathcal{M}_{\hat{I}}m/n + 2m' \mathcal{M}_{\hat{I}}W/n$ . By triangle inequality,  $\|\bar{\Omega} - \hat{\Omega}\| \leq \|W' \mathcal{P}_{\hat{I}}W/n\| + \|m' \mathcal{M}_{\hat{I}}m/n\| + \|2m' \mathcal{M}_{\hat{I}}W/n\|$ . Bounds for each of the three previous terms are established in Lemma 4 giving  $\|\bar{\Omega} - \hat{\Omega}\| = O_P(K\phi/n)$ .

Since  $\Omega$  has minimal eigenvalues bounded from below by assumption, it follows that  $\hat{\Omega}$  is invertible with probability approaching 1 (by  $\bar{\Omega}$  being invertible with probability approaching 1 and by Lemma 4). Consider the event  $\mathcal{L} = \{\lambda_{\min}(\hat{\Omega}) > \lambda_{\min}(\Omega)/2\}$ . By reasoning identical to that given in [29], the following "variance" and "bias" terms have bounds  $1_{\mathcal{L}}\|\bar{\Omega}^{-1}W'\epsilon/n\| = O_P(\sqrt{K}/\sqrt{n})$  and  $1_{\mathcal{L}}\|\bar{\Omega}^{-1}W'(G - P\beta)/n\| = O_P(K^{-\alpha})$ . To proceed, it is required to obtain analogous bounds for  $1_{\mathcal{L}}\|\hat{\Omega}^{-1}P' \mathcal{M}_{\hat{I}}\epsilon/n\|$  and  $1_{\mathcal{L}}\|\hat{\Omega}^{-1}P' \mathcal{M}_{\hat{I}}(G - P\beta)/n\|$ . Considering the "variance" term first, note that

$$\begin{aligned} 1_{\mathcal{L}}\|\hat{\Omega}^{-1}P' \mathcal{M}_{\hat{I}}\epsilon/n - \bar{\Omega}^{-1}W'\epsilon/n\| &\leq 1_{\mathcal{L}}\|(\hat{\Omega}^{-1} - \bar{\Omega}^{-1})W'\epsilon/n\| \\ &\quad + 1_{\mathcal{L}}\|\bar{\Omega}^{-1}(W' - P' \mathcal{M}_{\hat{I}})\epsilon/n\|. \end{aligned}$$

Consider the first term above.  $1_{\mathcal{L}}\|(\hat{\Omega}^{-1} - \bar{\Omega}^{-1})W'\epsilon/n\| \leq 1_{\mathcal{L}}\lambda_{\max}(\hat{\Omega}^{-1} - \bar{\Omega}^{-1})\|W'\epsilon/n\| = O_P(\sqrt{K}/n)O_P(\zeta_0(K)\sqrt{K}/\sqrt{n}) = O_P(\sqrt{K}/n)$ . For the second term,  $1_{\mathcal{L}}\|\bar{\Omega}^{-1}(W' - P' \mathcal{M}_{\hat{I}})\epsilon/n\| \leq 1_{\mathcal{L}}\lambda_{\max}(\bar{\Omega}^{-1})\|(W' - P' \mathcal{M}_{\hat{I}})\epsilon/n\| = 1_{\mathcal{L}}O_P(1)\|m' \mathcal{M}_{\hat{I}}\epsilon/n\| = O_P(\sqrt{K}\phi/\sqrt{n})$  by Lemma 4.

Turning to the "bias" term,

$$\begin{aligned} 1_{\mathcal{L}}\|\hat{\Omega}^{-1}P' \mathcal{M}_{\hat{I}}(G - P\beta)/n\| &= 1_{\mathcal{L}}[(G - P\beta)' \mathcal{M}_{\hat{I}}P\hat{\Omega}^{-1}P' \mathcal{M}_{\hat{I}}(G - P\beta)/n]^{1/2} \\ &= O_P(1)[(G - P\beta)'(G - P\beta)/n]^{1/2} \\ &= O_P(K^{-\alpha}) \end{aligned}$$

by assumption on  $(G - P\beta)$  and idempotency of  $\mathcal{M}_{\hat{I}}P\hat{\Omega}^{-1}P' \mathcal{M}_{\hat{I}} = \mathcal{M}_{\hat{I}}P(P' \mathcal{M}_{\hat{I}}P)^{-1} \mathcal{M}_{\hat{I}}$ .

The last intermediate ingredient before putting together the proof of the proposition is a bound on  $1_{\mathcal{L}}\|\hat{\Omega}^{-1}P' \mathcal{M}_{\hat{I}}H/n\| = O_P(1)\|P' \mathcal{M}_{\hat{I}}G_2/n\| = O_P(\zeta_0(K)\sqrt{K\phi/n} + \sqrt{K\phi^2/ns} + \sqrt{K}K^{-\alpha}\sqrt{\phi/s})/\sqrt{n} + O_P(\sqrt{K\phi}K^{-\alpha} + \sqrt{K}\sqrt{K^2\phi^2/n}\sqrt{\phi/n})/\sqrt{n}$  by triangle inequality and Lemma 4(iv) and 4(v). This reduces to  $O_P(\sqrt{K/n} + K^{-\alpha})$ .

To show the proposition, bound the difference  $\hat{\beta} - \beta$ . Note that  $1_{\mathcal{L}}(\hat{\beta} - \beta) = 1_{\mathcal{L}}\hat{\Omega}^{-1}P' \mathcal{M}_{\hat{I}}\epsilon/n + 1_{\mathcal{L}}\hat{\Omega}^{-1}P' \mathcal{M}_{\hat{I}}(G - P\beta)/n + 1_{\mathcal{L}}\hat{\Omega}^{-1}P' \mathcal{M}_{\hat{I}}H/n$ . Triangle inequality and bounds described above give  $1_{\mathcal{L}}\|\hat{\beta} - \beta\| \leq$

$1_{\mathcal{L}}\|\widehat{\Omega}^{-1}P'\mathcal{M}_{\widehat{\Gamma}}\epsilon/n\| + 1_{\mathcal{L}}\|\widehat{\Omega}^{-1}P'\mathcal{M}_{\widehat{\Gamma}}(G - P\beta)/n\| + 1_{\mathcal{L}}\|\widehat{\Omega}^{-1}P'\mathcal{M}_{\widehat{\Gamma}}H/n\| = O_p(K^{1/2}/\sqrt{n} + K^{-\alpha})$ . The statement of the proposition follows from the bound on  $\widehat{\beta} - \beta$  using the arguments in [29].

**B.3. Proof of Theorem.** Let  $F = V^{-1/2}$  and  $\bar{g} = p(x)'\beta$  and decompose the quantity  $1_{\mathcal{L}}\sqrt{n}F[a(\widehat{g}) - a(g)]$  by

$$1_{\mathcal{L}}\sqrt{n}F[a(\widehat{g}) - a(g)] = 1_n\sqrt{n}F[a(\widehat{g}) - a(g) + D(\widehat{g}) - D(g) + D(\bar{g}) - D(g) + D(\widehat{g}) - D(\bar{g})].$$

By arguments given in the proof of Theorem 2 in Newey (1997),  $1_{\mathcal{L}}|\sqrt{n}F[D(\bar{g}) - D(g)]| \leq C\sqrt{n}K^{-\alpha}$ . In addition, bounds on  $|\widehat{g} - g|_d$  given by the proposition imply that  $|\sqrt{n}F[a(\widehat{g}) - a(g) - D(\widehat{g}) + D(g)]| \leq C_{Lip}\sqrt{n}|\widehat{g} - g|_d^2 = O_P(C_{Lip}\sqrt{n}\zeta_d(K)(\sqrt{K}/\sqrt{n} + K^{-\alpha} + \sqrt{s}/\sqrt{n})^2) \rightarrow 0$ . It remains to be shown that  $1_{\mathcal{L}}\sqrt{n}F[D(\widehat{g}) - D(\bar{g})]$  satisfies an appropriate central limit theorem. Note that  $D(\widehat{g})$  can be expanded

$$\begin{aligned} D(\widehat{g}) &= D(p(x)'\widehat{\beta}) = D(p(x)'\widehat{\Omega}^{-1}P'\mathcal{M}_{\widehat{\Gamma}}y) \\ &= D(p(x)'\widehat{\Omega}^{-1}P'\mathcal{M}_{\widehat{\Gamma}}(G + H + \epsilon)) = D(p(x)'\widehat{\Omega}^{-1}P'\mathcal{M}_{\widehat{\Gamma}}(G + H + \epsilon)) \\ &= A'\widehat{\Omega}^{-1}P'\mathcal{M}_{\widehat{\Gamma}}(G + H + \epsilon) = A'\widehat{\Omega}^{-1}P'\mathcal{M}_{\widehat{\Gamma}}G + A'\widehat{\Omega}^{-1}P'\mathcal{M}_{\widehat{\Gamma}}H + A'\widehat{\Omega}^{-1}P'\mathcal{M}_{\widehat{\Gamma}}\epsilon \end{aligned}$$

Using the above expansion and  $D(\bar{g}) = D(p(x)'\beta) = A'\beta$  gives

$$\begin{aligned} \sqrt{n}F[D(\widehat{g}) - D(\bar{g})] &= \sqrt{n}FA'[\widehat{\Omega}^{-1}P'\mathcal{M}_{\widehat{\Gamma}}G - \beta] \\ &\quad + \sqrt{n}FA'[\widehat{\Omega}^{-1}P'\mathcal{M}_{\widehat{\Gamma}}H] + \sqrt{n}FA'[\widehat{\Omega}^{-1}P'\mathcal{M}_{\widehat{\Gamma}}\epsilon] \end{aligned}$$

The terms  $\sqrt{n}FA'[\widehat{\Omega}^{-1}P'\mathcal{M}_{\widehat{\Gamma}}G - \beta]$  and  $\sqrt{n}FA'[\widehat{\Omega}^{-1}P'\mathcal{M}_{\widehat{\Gamma}}H]$  will be shown negligible while the third term  $\sqrt{n}FA'[\widehat{\Omega}^{-1}P'\mathcal{M}_{\widehat{\Gamma}}\epsilon]$  will be shown satisfying a central limit theorem.

First, note the expressions  $1_{\mathcal{L}}\|FA'\widehat{\Omega}^{-1}\| = O_P(1)$ ,  $1_{\mathcal{L}}\|FA'\widehat{\Omega}^{-1/2}\| = O_P(1)$  both hold by arguments in Newey (1997). Beginning with the first term,

$$\begin{aligned} &1_{\mathcal{L}}|\sqrt{n}FA'[\widehat{\Omega}^{-1}P'\mathcal{M}_{\widehat{\Gamma}}G/n - \beta]| \\ &= 1_{\mathcal{L}}|\sqrt{n}FA'[(P'\mathcal{M}_{\widehat{\Gamma}}P/n)^{-1}P'\mathcal{M}_{\widehat{\Gamma}}(G - P\beta)/n]| \\ &\leq 1_{\mathcal{L}}\|FA'\widehat{\Omega}^{-1}P'\mathcal{M}_{\widehat{\Gamma}}/\sqrt{n}\|\|G - P\beta\| \\ &\leq 1_{\mathcal{L}}\|FA'\widehat{\Omega}^{-1/2}\|\sqrt{n}\max_{i \leq n}|g(x_i) - \bar{g}(x_i)| \\ &\leq 1_{\mathcal{L}}\|FA'\widehat{\Omega}^{-1/2}\|\sqrt{n}\|g - \bar{g}\|_0 = O_P(1)O_P(\sqrt{n}K^{-\alpha}) = o_P(1) \end{aligned}$$

Next, consider  $\sqrt{n}FA'\widehat{\Omega}^{-1}P'\mathcal{M}_{\widehat{\Gamma}}H/n$ . By  $P = m + W$ , triangle inequality, Cauchy-Schwartz and Lemma 4,

$$\begin{aligned} |FA'\widehat{\Omega}^{-1}P'\mathcal{M}_{\widehat{\Gamma}}H/\sqrt{n}| &\leq |FA'\widehat{\Omega}^{-1}m'\mathcal{M}_{\widehat{\Gamma}}G_2/\sqrt{n}| + |FA'\widehat{\Omega}^{-1}W'\mathcal{M}_{\widehat{\Gamma}}H/\sqrt{n}| \\ &\leq \|FA'\widehat{\Omega}^{-1}\|\|m'\mathcal{M}_{\widehat{\Gamma}}H/\sqrt{n}\| + \|FA'\widehat{\Omega}^{-1}\|\|W'\mathcal{M}_{\widehat{\Gamma}}H/\sqrt{n}\| \\ &= O_P(1)o_P(1) + O_P(1)o_P(1) \end{aligned}$$

Next consider the last remaining term for which a central limit result will be shown. Note that using the bounds for  $\|P' \mathcal{M}_{\hat{\Gamma}} \epsilon / \sqrt{n}\|$  and  $\|m' \mathcal{M}_{\hat{\Gamma}} \epsilon / \sqrt{n}\|$  derived in Lemma 4,

$$\begin{aligned}
\sqrt{n} F A' \hat{\Omega}^{-1} P' \mathcal{M}_{\hat{\Gamma}} \epsilon / n &= \sqrt{n} F A' \Omega^{-1} P' \mathcal{M}_{\hat{\Gamma}} \epsilon / n + \sqrt{n} F A' (\hat{\Omega}^{-1} - \Omega^{-1}) P' \mathcal{M}_{\hat{\Gamma}} \epsilon / n \\
&= \sqrt{n} F A' \Omega^{-1} P' \mathcal{M}_{\hat{\Gamma}} \epsilon / n + O(\|F A' (\hat{\Omega}^{-1} - \Omega^{-1})\| \|P' \mathcal{M}_{\hat{\Gamma}} \epsilon / \sqrt{n}\|) \\
&= \sqrt{n} F A' \Omega^{-1} P' \mathcal{M}_{\hat{\Gamma}} \epsilon / n + o_P(1) \\
&= \sqrt{n} F A' \Omega^{-1} W' \epsilon / n + \sqrt{n} F A' \Omega^{-1} (W' - P' \mathcal{M}_{\hat{\Gamma}}) \epsilon / n + o_P(1) \\
&= \sqrt{n} F A' \Omega^{-1} W' \epsilon / n + O(\|F A' \Omega^{-1}\| \|m' \mathcal{M}_{\hat{\Gamma}} \epsilon / \sqrt{n}\|) + o_P(1) \\
&= \sqrt{n} F A' \Omega^{-1} W' \epsilon / n + o_P(1)
\end{aligned}$$

Let  $Z_{in} = F A' W_i \epsilon_i / \sqrt{n}$ . Then  $\sum_i Z_{in} = F A' V' \epsilon / \sqrt{n}$ . For each  $n$ ,  $Z_{in}$  is i.i.d. with  $E[Z_{in}] = 0$ ,  $\sum_i E[Z_{in}^2] = 1$ . In addition,

$$\begin{aligned}
n E[1_{\{|Z_{in}| > \delta\}} Z_{in}^2] &= n \delta^2 E[1_{\{|Z_{in}|/\delta > 1\}} Z_{in}^2 / \delta^2] \leq n \delta^2 E[Z_{in}^4 / \delta^4] \\
&\leq n \delta^2 \|F A'\|^4 \zeta_0(K)^2 E[\|w_i\|^2 E[\epsilon_i^4 | x_i]] / n^2 \delta^4 \leq C \zeta_0(K)^2 K_1 / n \rightarrow 0.
\end{aligned}$$

By the Lindbergh-Feller Central Limit Theorem,  $\sum_i Z_{in} \xrightarrow{d} N(0, 1)$ .

Next consider the plug in variance estimate. First, bound  $\|\hat{A} - A\|$ . In the case that  $a(g)$  is linear in  $g$ , then  $a(p' \beta) = A' \beta \implies \hat{A} = A$ . Therefore, it is sufficient to consider the case (ii) of Assumption 7, that  $a(g)$  is not linear in  $g$ . For  $\nu$  as in the statement of Assumption 7, Define the event  $\mathcal{E} = \mathcal{E}_n = \{|\hat{g} - g|_d < \nu/2\}$ . In addition, let  $\hat{J} = (D(p_{11}; \hat{g}), \dots, D(p_{1K}; \hat{g}))'$ . Then for any  $\beta$  such that  $|p' \beta - \hat{g}| < \nu/2$ , it follows that  $|p' \beta - g| \leq \nu$  and

$$\begin{aligned}
&1_{\mathcal{E}} |a(p' \beta) - a(\hat{g}) - \hat{J}'(\beta - \hat{\beta})| / \|\beta - \hat{\beta}\| \\
&= 1_{\mathcal{E}} |a(p' \beta) - a(\hat{g}) - D(p' \beta; \hat{g}) + D(\hat{g}; \hat{g})| / \|\beta - \hat{\beta}\| \\
&\leq 1_{\mathcal{E}} C \cdot |p' \beta - \hat{g}|_d^2 / \|\beta - \hat{\beta}\| \leq 1_{\mathcal{E}} C \cdot \zeta_d(K)^2 \|\beta - \hat{\beta}\| \rightarrow 0
\end{aligned}$$

Therefore,  $\hat{A}$  exists and equals  $\hat{J}$  if  $1_{\mathcal{E}} = 1$ .

$$\begin{aligned}
1_{\mathcal{E}} \|\hat{A} - A\|^2 &= 1_{\mathcal{E}} (\hat{A} - A)' (\hat{A} - A) = 1_{\mathcal{E}} |D((\hat{A} - A)' p; \hat{g}) - D((\hat{A} - A)' p; g)| \\
&\leq C \cdot 1_{\mathcal{E}} |(\hat{A} - A)' p|_d |\hat{g} - g|_d \leq C \cdot \|\hat{A} - A\| \zeta_d(K) |\hat{g} - g|_d
\end{aligned}$$

This gives  $1_{\mathcal{E}} \|\hat{A} - A\| \leq C \cdot \zeta_d(K) |\hat{g} - g|_d = O_P(\zeta_d(K)^2 (\sqrt{K}/\sqrt{n} K^{-\alpha})) \xrightarrow{P} 0$ .

A consequence of the bound on  $\|\hat{A} - A\|$  is that  $1_{\mathcal{E}} \|F \hat{A}\| \leq 1_{\mathcal{E}} \|F\| \|\hat{A} - A\| + \|F A\| = O_P(1)$ . Similarly,  $1_{\mathcal{E}} \|F \hat{A} \hat{\Omega}^{-1}\| = O_P(1)$ . Next, define  $\hat{u} = 1_{\mathcal{E}} \hat{\Omega}^{-1} \hat{A} F$  and  $u = 1_{\mathcal{E}} \Omega^{-1} A F$ .

$$\begin{aligned}
\|\hat{u} - u\| &\leq 1_{\mathcal{E}} \|F \hat{A}' \hat{\Omega}^{-1} (\Omega - \hat{\Omega})\| + 1_{\mathcal{E}} \|F (\hat{A} - A)'\| \\
&\leq 1_{\mathcal{E}} \|F \hat{A}' \hat{\Omega}^{-1}\| \|\Omega - \hat{\Omega}\| + 1_{\mathcal{E}} \|F\| \|\hat{A} - A\| \xrightarrow{P} 0
\end{aligned}$$



Next, note that  $u'\Sigma u = 1_{\mathcal{E}}$ . In addition,  $\Sigma \leq C \cdot I$  in the positive definite sense by Assumption. Therefore,

$$\begin{aligned} 1_{\mathcal{E}}|\hat{u}'\Sigma\hat{u} - 1| &= |\hat{u}\Sigma\hat{u} - u'\Sigma u| \leq (\hat{u} - u)'\Sigma(\hat{u} - u) + |2(\hat{u} - u)'\Sigma u| \\ &\leq C \cdot \|\hat{u} - u\|^2 + 2((\hat{u} - u)'\Sigma(\hat{u} - u))^{1/2}(u'\Sigma u)^{1/2} \\ &\leq o_P(1) + C\|\hat{u} - u\| \xrightarrow{P} 0. \end{aligned}$$

Define  $\tilde{\Sigma} = \sum_i W_i W_i' \epsilon_i^2 / n$ , an infeasible sample analogue of  $\Sigma$ . By reasoning similar to that showing  $\|\tilde{\Omega} - \Omega\| \xrightarrow{P} 0$  it follows that  $\|\tilde{\Sigma} - \Sigma\| \xrightarrow{P} 0$ . Then this implies that  $1_{\mathcal{E}}|\hat{u}\tilde{\Sigma}\hat{u} - \hat{u}'\Sigma\hat{u}| = |\hat{u}'(\tilde{\Sigma} - \Sigma)\hat{u}| \leq \|\hat{u}\|^2 \|\tilde{\Sigma} - \Sigma\| = O_P(1)o_P(1) \xrightarrow{P} 0$ .

Next, let  $\Delta_{1i} = g(x_i) - \hat{g}(x_i)$  and  $\Delta_{2i} = h(z_i) - \hat{h}(z_i)$ . Then  $\max_{i \leq n} |\Delta_i| \leq |\hat{g} - g|_0 = o_P(1) \xrightarrow{P} 0$  follows from the proposition above. Let  $\omega_i^2 = \hat{u}'W_i W_i' \hat{u}$  and  $\hat{\omega}_i^2 = \hat{u}'W_i W_i' \hat{u}$ . Bound  $\tilde{\Sigma}$  to  $\hat{\Sigma}$  by considering the quantity

$$\begin{aligned} \mathcal{E}_n|F\hat{V}F - \hat{u}'\tilde{\Sigma}\hat{u}| &= |\hat{u}'(\hat{\Sigma} - \tilde{\Sigma})\hat{u}| = \left| \sum_{i=1}^n \hat{u}'\hat{W}_i \hat{W}_i' \hat{\epsilon}_i^2 \hat{u} / n - \sum_{i=1}^n \hat{u}'W_i W_i' \epsilon_i^2 \hat{u} / n \right| \\ &\leq \left| \sum_{i=1}^n \omega_i^2 (\hat{\epsilon}_i^2 - \epsilon_i^2) / n \right| + \left| \sum_{i=1}^n (\hat{\omega}_i^2 - \omega_i^2) \epsilon_i^2 / n \right| \end{aligned}$$

Both terms on the right hand side will be bounded. Consider the first term. Expanding  $(\hat{\epsilon}_i^2 - \epsilon_i^2)$  gives

$$\begin{aligned} \left| \sum_{i=1}^n \omega_i^2 (\hat{\epsilon}_i^2 - \epsilon_i^2) / n \right| &\leq \left| \sum_{i=1}^n \omega_i^2 \Delta_{1i}^2 / n \right| + \left| \sum_{i=1}^n \omega_i^2 \Delta_{2i}^2 / n \right| + \left| \sum_{i=1}^n \omega_i^2 \Delta_{1i} \Delta_{2i} / n \right| \\ &\quad + 2 \left| \sum_{i=1}^n \omega_i^2 \Delta_{1i} \epsilon_i / n \right| + 2 \left| \sum_{i=1}^n \omega_i^2 \Delta_{2i} \epsilon_i / n \right| \end{aligned}$$

These five terms above are bounded in order of their appearance by

$$\begin{aligned}
\sum_{i=1}^n \omega_i^2 \Delta_{1i}^2 / n &\leq \max_{i \leq n} |\Delta_{1i}| \sum_{i=1}^n \omega_i^2 / n = o_P(1) O_P(1) \\
\sum_{i=1}^n \omega_i^2 \Delta_{2i}^2 / n &\leq \max_{i \leq n} |\Delta_{2i}| \sum_{i=1}^n \omega_i^2 |\epsilon_i| / n = o_P(1) O_P(1) \\
\sum_{i=1}^n \omega_i^2 \Delta_{1i} \Delta_{2i} / n &\leq \max_{i \leq n} |\Delta_{1i}| \max_{i \leq n} |\Delta_{2i}| \sum_{i=1}^n \omega_i^2 / n = o_P(1) O_P(1) \\
\sum_{i=1}^n \omega_i^2 \Delta_{1i} \epsilon_i / n &\leq \max_{i \leq n} |\Delta_{1i}| \sum_{i=1}^n \omega_i^2 |\epsilon_i| / n = o_P(1) O_P(1) \\
\sum_{i=1}^n \omega_i^2 \Delta_{2i} \epsilon_i / n &\leq \max_{i \leq n} |\Delta_{2i}| \sum_{i=1}^n \omega_i^2 |\epsilon_i| / n = o_P(1) O_P(1)
\end{aligned}$$

Where the bounds  $\max_{i \leq n} |\Delta_{1i}| = o_P(1)$  follows by the proposition and  $\max_{i \leq n} |\Delta_{2i}| = o_P(1)$  follows from Lemma 5 below. On the other hand, the argument that  $\sum_{i=1}^n \omega_i^2 / n, \sum_{i=1}^n \omega_i^2 |\epsilon_i| = O_P(1)$  is the same as in [29].

The second term is bounded by

$$\begin{aligned}
&\left| \sum_{i=1}^n \hat{u}(\widehat{W}_i \widehat{W}_i' - W_i W_i') \hat{\epsilon}_i^2 \hat{u} / n \right| \leq \max_{i \leq n} |\epsilon_i^2| \left| \sum_{i=1}^n \hat{u}(\widehat{W}_i \widehat{W}_i' - W_i W_i') \hat{u} / n \right| \\
&\leq \max_{i \leq n} |\epsilon_i^2| \|\hat{u}\|^2 \sum_{i=1}^n (\widehat{W}_i \widehat{W}_i' - W_i W_i') / n = \max_{i \leq n} |\epsilon_i^2| \|\hat{u}\|^2 \|\widehat{\Omega} - \bar{\Omega}\| \\
&\leq \left( \max_{i \leq n} |\epsilon_i^2| + \max_{i \leq n} |\hat{\epsilon}_i^2 - \epsilon_i^2| \right) \|\hat{u}\|^2 \|\widehat{\Omega} - \bar{\Omega}\| \\
&= \left( O_P(n^{2/q}) + o_P(1) \right) O_P(1) O_P(K\phi/\sqrt{n}) \\
&= o_P(1)
\end{aligned}$$

where the last bounds come from the rate condition in Assumption 9 and  $\max_{i \leq n} |\hat{\epsilon}_i^2 - \epsilon_i^2| = o_P(1)$  by  $\max_{i \leq n} |\Delta_{1i}| + |\Delta_{2i}| = o_P(1)$ .

This implies that  $\mathcal{E}_n |F\widehat{V}F - 1| \xrightarrow{P} 0$ . With probability approaching 1,  $1_{\mathcal{E}} = 1$ , this gives  $F\widehat{V}F \xrightarrow{P} 1$  which in turn implies that

$$\sqrt{n}\widehat{V}^{-1/2}(\widehat{\theta} - \theta) = \sqrt{n}F(\widehat{\theta} - \theta)/(F\widehat{V}F)^{1/2} \xrightarrow{d} N(0, 1).$$

To provide a rate of convergence,  $|V| \leq C \cdot \zeta_d(K)^2$  since  $\widehat{\theta} = \theta_0 + (V^{1/2}/\sqrt{n})\sqrt{n}F(\widehat{\theta} - \theta) = \theta + O_P(V^{1/2}/\sqrt{n})$ . Cauchy-Schwartz inequality implies that  $|p'\beta|_d \leq \zeta_d(K)\|\beta\|$  for any choice of  $\beta$ . Then  $\|A\|^2 = |D(p'A)| \leq C \cdot |p'A|_d \leq C \cdot \zeta_d(K)\|A\|$ . This gives  $\|A\| \leq C \cdot \zeta_d(K)$  and  $|V| \leq C \cdot \|A\|^2 \leq C \cdot \zeta_d(K)^2$ .

The proof of the second statement of the Theorem uses similar arguments as the proof of the first and follows from the proof of Theorem 3 in [29].

### APPENDIX C. LEMMAS

The first lemma is a performance bound for Post-Lasso estimates. It is required for use in the next two Lemmas. It is based on the results of [3]. Define the following four events which are useful for describing the regularization properties of the lasso regressions.

$$\mathcal{A}_{FS} = \{\lambda^{FS}/n \geq c \max_{j \leq L, k \leq K} |S_{k,j}|\}, \quad \mathcal{A}_{RF} = \{\lambda^{RF}/n \geq c \max_{j \leq L} |S_{j,RF}|\}$$

$$\mathcal{B}_{FS} = \{\ell \Psi_{jk}^{FS} \leq \widehat{\Psi}_{jk}^{FS} \leq u \Psi_{ij}^{FS} \forall j, k\}, \quad \mathcal{B}_{RF} = \{\ell \Psi_j^{RF} \leq \widehat{\Psi}_j^{RF} \leq u \Psi_j^{RF} \forall j\}.$$

Where  $\ell$  and  $u$  are the constants in Assumption 6 and  $S_{j,RF}, S_{k,j}$  are defined as the scores of the quadratic components of the Lasso optimization problems in Section 3.3. Define the regularization event  $\mathcal{R} = \mathcal{A}_{FS} \cap \mathcal{A}_{RF} \cap \mathcal{B}_{FS} \cap \mathcal{B}_{RF}$ . In addition, define  $c_0 = (uc + 1)/(\ell c - 1)$ . Let  $\kappa_C = \min_{\delta \in \Delta_{C,\mathcal{T}} | \mathcal{T}| \leq s} s \delta' (Q'Q/n) \delta / \|\delta_{\mathcal{T}}\|_1^2$  where  $\Delta_{C,\mathcal{T}} = \{\delta \neq 0 : \|\delta_{\mathcal{T}^c}\|_1 \leq C \|\delta_{\mathcal{T}}\|_1\}$ . This defines the restricted eigenvalue and is useful for Lasso bounds. For more details regarding the definition, see for example, [10]. Let  $\kappa_{c_0}^k = \min_{\|\Psi_k^{FS} \delta_{\mathcal{T}_k^c}\|_1 \leq c_0 \|\Psi_k^{FS} \delta_{\mathcal{T}_k}\|_1} \sqrt{s} \delta' (Q'Q/n) \delta / \|\widehat{\Psi}_k^{FS} \delta_{\mathcal{T}_k}\|_1$ . Define  $\kappa_{c_0}^{RF}$  analogously using the reduced form  $\widehat{\Psi}^{RF}$ .

**Lemma 1.** *Under the conditions given in the text, the following inequalities holds.*

$$1_{\mathcal{R}} \max_{k \leq K} \|\mathcal{M}_{\widehat{I}} m_k / \sqrt{n}\|_2 \leq 1_{\mathcal{R}} \left( \max_{k \leq K} (u + 1/c) \frac{\lambda^{FS} \sqrt{s}}{n \kappa_{c_0}^k} + 3\xi_{FS} \right).$$

$$1_{\mathcal{R}} \|\mathcal{M}_{\widehat{I}} E[y|z] / \sqrt{n}\|_2 \leq 1_{\mathcal{R}} \left( (u + 1/c) \frac{\lambda^{RF} \sqrt{s}}{n \kappa_{c_0}^{RF}} + 3\xi_{RF} \right).$$

In addition, the regularization event satisfies  $P(\mathcal{R}) \rightarrow 1$ .

*Proof.* That  $\mathcal{A}_{RF}$  holds with probability approaching 1 was established in [3]. The conditions listed in Assumption 6 allow use of the same argument to show that  $\mathcal{A}_{FS}$  holds with high probability by allowing the application of the moderate deviation results of [22]. In addition,  $\mathcal{B}_{FS}, \mathcal{B}_{RF}$  hold by Assumption 6.

Therefore,  $1_{\mathcal{R}} \xrightarrow{P} 1$  giving the last claim of the lemma. The first two claims follow immediately from the third statement of Lemma 7 in [3].  $\square$

**Lemma 2.**  $\max_{k \leq K} (1/\kappa_{c_0}^k) = O_P(1)$  and  $\max_{k \leq K} |\{j : \widehat{\Gamma}_{kj} \neq 0\}| = O_P(s)$

*Proof.* For the first result, let  $a = \min_{k,j} |\widehat{\Psi}_{kj}^{FS}|$  and  $b = \max_{k,j} |\widehat{\Psi}_{kj}^{FS}|$ . Step 1 of the the proof of Theorem 1 in [3] shows that  $\max_{k \leq K} (1/\kappa_{c_0}^k) \leq b(\kappa_{bc_0/a}(Q'Q/n))^{-1}$ . As a simple consequence of Assumption 6,  $a$  and  $b$

are bounded from above and away from zero with probability approaching 1. Then, also by assumption,  $(\kappa_{bc_0/a}(Q'Q/n))^{-1} = O_P(1)$ . This implies the first statement. For the second statement, let  $\hat{s}_k$  be the number of incorrectly selected terms in the  $k$ -th first stage regression. Then By Lemma 10 of [3],  $\hat{s}_k \leq s\varphi_{\max}(q) \max_j |\Psi_{jk}|^4 (2c_0/\kappa_{c_0}^k + 6c_0n\xi_{FS}/\lambda\sqrt{s})^2$  for every integer  $q > 2s\varphi_{\max}(q) \max_j |\Psi_{jk}|^{-2} (2c_0/\kappa_{c_0}^k + 6c_0n\xi_{FS}/\lambda\sqrt{s})^2$ . The choice  $q = \kappa'' 2s\varphi_{\max}(q) \max_{k,j} |\Psi_{jk}|^{-2} (2c_0/\kappa_{c_0}^k + 6c_0n\xi_{FS}/\lambda\sqrt{s})^2$  yields  $\varphi_{\max}(q) = O_P(1)$  by Assumption 5 and by using  $\max_{k,j} |\Psi_{jk}|^{-2} = O_P(1)$ ,  $\max_{j,k} |\Psi_{jk}|^4 = O_P(1)$ ,  $\max_k 2c_0/\kappa_{c_0}^k = O_P(1)$  and  $6c_0n\xi_{FS}/\lambda^{FS}\sqrt{s} = O_P(1)$  it follows that  $\max_{k \leq K} \hat{s}_k = O_P(Ks)$ .  $\square$

The following lemmas bounds various quantities used in the proof above. The lemma provides analogous results to steps 4-6 of [9] in the proof of their Theorem 1 but accounts increasing number of series terms in the first dictionary.

**Lemma 3.** *First Stage and Reduced Form Performance Bounds*

- (i)  $\max_{k \leq K} \|\mathcal{M}_{\hat{I}} m_k / \sqrt{n}\| = O_P(\sqrt{\phi/n})$
- (ii)  $\|\mathcal{M}_{\hat{I}} H / \sqrt{n}\| = O_P(\sqrt{K\phi/n} + K^{-\alpha})$
- (iii)  $\max_{k \leq K} \|\hat{\Gamma}_k(\hat{I}) - \Gamma_k\| = O_P(\sqrt{\phi/n})$
- (iv)  $\|b(H; \hat{I}) - \eta\| = O_P(\sqrt{\phi/n} + K^{-\alpha})$
- (v)  $\max_{k \leq K} \|Q'W_k / \sqrt{n}\|_{\infty} = O_P(\sqrt{\phi/s}), \|Q'\epsilon / \sqrt{n}\|_{\infty} = O_P(\sqrt{\phi/s})$ .
- (vi)  $\max_{k \leq K} \|b(W_k; \hat{I})\|_1 = O_P(\sqrt{K^2 s \phi/n})$

*Proof.* Statement (i) follows from an application of Lemma 1:

$$\begin{aligned}
1_{\mathcal{R}} \max_{k \leq K} \|\mathcal{M}_{\hat{I}} m_k / \sqrt{n}\| &\leq 1_{\mathcal{R}} \max_{k \leq K} \|\mathcal{M}_{\hat{I}_k} m_k / \sqrt{n}\| \\
&\leq 1_{\mathcal{R}} \max_{k \leq K} (u + 1/c) \lambda^{FS} \sqrt{s/n} \kappa_{c_0}^k + 3\xi_{FS} \\
&\leq 1_{\mathcal{R}} \max_{k \leq K} (u + 1/c) (C \sqrt{n \log(\max(KL, n))}) \sqrt{s/n} \kappa_{c_0}^k + 1_{\mathcal{R}} 3\xi_{FS} \\
&= O(\sqrt{\phi/n}) / \kappa_{c_0}^k + 3\xi_{FS} = O_P(\sqrt{\phi/n})
\end{aligned}$$

Where the last equality follows from Lemma 2 and the definition of  $\lambda^{FS}$ ,  $\xi_{FS} = O_P(\sqrt{\phi/n})$  and  $1_{\mathcal{R}} \xrightarrow{P} 1$ . Next, Consider statement (ii).

$$\begin{aligned}
1_{\mathcal{R}} \|\mathcal{M}_{\hat{I}} H / \sqrt{n}\| &\leq 1_{\mathcal{R}} \|\mathcal{M}_{\hat{I}} (E[G|z] + H) / \sqrt{n}\| + 1_{\mathcal{R}} \|\mathcal{M}_{\hat{I}} E[G|z] / \sqrt{n}\| \\
&\leq 1_{\mathcal{R}} \|\mathcal{M}_{I_{RF}} (E[y|z]) / \sqrt{n}\| + 1_{\mathcal{R}} \|\mathcal{M}_{\hat{I}} E[G|z] / \sqrt{n}\| \\
&\leq 1_{\mathcal{R}} (u + 1/c) \lambda^{RF} \sqrt{Ks/n} \kappa_{c_0}^{RF} + 1_{\mathcal{R}} 3\xi_{RF} + 1_{\mathcal{R}} \|\mathcal{M}_{\hat{I}} E[G|z] / \sqrt{n}\| \\
&\leq 1_{\mathcal{R}} (u + 1/c) (C \sqrt{n \log(\max(KL, n))}) \sqrt{Ks/n} \kappa_{c_0}^{RF} \\
&\quad + 1_{\mathcal{R}} 3\xi_{RF} + 1_{\mathcal{R}} \|\mathcal{M}_{\hat{I}} E[G|z] / \sqrt{n}\|
\end{aligned}$$

To control the approximation error for the reduced form, ie. to bound  $\xi_{RF}$ , note that

$$g(x) + h(z) = q(z)'(\Gamma\beta + \eta) + (g(x) - p(x)\beta) + (h(z) - q(z)'\eta)$$

The approximation error  $\xi_{RF}$  is then given by

$$\begin{aligned}\xi_{RF}^2 &= (G - P\beta + H - Q\eta)'(G - P\beta + H - Q\eta)/n \\ &\leq 2(G - P\beta)'(G - P\beta)/n + 2(H - Q\eta)'(H - Q\eta)/n \\ &= O(K^{-2\alpha}) + O(Ks/n)\end{aligned}$$

Next consider  $\|\mathcal{M}_{\hat{T}}E[G|z]/\sqrt{n}\|$  and note that  $E[G|z] = m\beta + E[G - m\beta|z]$ . By statement (i) of this lemma,  $\|\mathcal{M}_{\hat{T}}m\beta_1/\sqrt{n}\| \leq \max_k \|\mathcal{M}_{\hat{T}}m_k/\sqrt{n}\| \|\beta_1\| = O_P(\sqrt{\phi/n})O(1)$ . Next,  $\|\mathcal{M}_{\hat{T}}E[G - m\beta|z]/\sqrt{n}\| \leq \|\mathcal{M}_{\hat{T}}E[G - P\beta|z]/\sqrt{n}\| + \|\mathcal{M}_{\hat{T}}E[P\beta - m\beta|z]/\sqrt{n}\|$ . The first term  $\|\mathcal{M}_{\hat{T}}E[G - P\beta|z]/\sqrt{n}\|$  is  $O(K^{-\alpha})$  and the second term  $\|\mathcal{M}_{\hat{T}}E[P\beta - m\beta|z]/\sqrt{n}\|$  vanishes identically. These put together establish that  $1_{\mathcal{A}}\|\mathcal{M}_{\hat{T}}H/\sqrt{n}\| \leq 1_{\mathcal{A}}O_P(\sqrt{K\phi/n}) + O(K^{-\alpha})$ .

The result follows by noting that  $1_{\mathcal{A}} \xrightarrow{P} 1$ .

Next consider statement (iii). Let  $\hat{T} = \hat{I} \cup \text{supp}(\Gamma_1) \cup \dots \cup \text{supp}(\Gamma_K)$ .

$$\begin{aligned}\max_{k \leq K} \|\hat{\Gamma}_k(\hat{T}) - \Gamma_k\| &\leq \max_k \{\sqrt{\varphi_{\min}(|\hat{T}_k|)} \|\hat{\Gamma}_k(\hat{T}) - \Gamma_k\|\} \leq \max_{k \leq K} \|Q(\hat{\Gamma}_k(\hat{T}) - \Gamma_k)/\sqrt{n}\| \\ &\leq \max_{k \leq K} \{\|\mathcal{M}_{\hat{T}}m_k/\sqrt{n}\| + \|(m_k - Q\Gamma_k)/\sqrt{n}\|\} = O_P(\sqrt{\phi/n}).\end{aligned}$$

Where the last bound follows from  $\varphi_{\min}(\hat{T}) = O_P(1)$  by Assumption 5 on the restricted eigenvalues and by  $\hat{T} = O_P(Ks)$  by the result of the lemma above. Statement (iv) follows from similar reasoning as for statement (iii).

Statement (v): note that by Lemma 4 of [9], a sufficient condition for

$$\max_{k \leq K, j \leq L} \frac{|Q'_j W_k|/\sqrt{n}}{\sqrt{\sum_i q_j(z_i)^2 W_{ki}^2/n}} = O_P(\phi/s)$$

is that  $\min_{k \leq K, j \leq L} E[q_j(z_i)^2 W_{ki}^2]^{1/2} E[|q_j(z_i)|^3 |W_{ki}|^3]^{-1/3} = O(1)$  and  $\log(KL) = o(n^{1/3})$ . These conditions follow from Assumption 6. In addition,  $\sqrt{\sum_i q_j(z_i)^2 W_{ki}^2/n} = O_P(1)$  by Assumption 6. This gives the first part of statement (vi). The second part follows in the same manner.

Statement (vi):

$$\begin{aligned}\max_{k \leq K} \|b(W_k; \hat{T})\|_1 &\leq \max_{k \leq K} \sqrt{|\hat{T}|} \|b(W_k; \hat{T})\| \leq \max_{k \leq K} \sqrt{|\hat{T}|} (Q(\hat{T})' Q(\hat{T})/n)^{-1} Q(\hat{T})' W_k/n \\ &\leq \sqrt{|\hat{T}|} \varphi_{\min}^{-1}(|\hat{T}|) \sqrt{|\hat{T}|} \max_{k \leq K} \|Q' W_k/\sqrt{n}\|_{\infty} / \sqrt{n} = O_P(\sqrt{K^2 s \phi/n})\end{aligned}$$

□

**Lemma 4.** *The following bounds hold.*

- (i)  $\|W' \mathcal{P}_{\hat{T}} W/n\| = O_P(\sqrt{K^2/n} \sqrt{\phi^2/n})$
- (ii)  $\|m' \mathcal{M}_{\hat{T}} m/n\| = O_P(K\phi/n)$

$$\begin{aligned}
(iii) \quad & \|m' \mathcal{M}_{\hat{I}} W/n\| = O_P(\sqrt{K^2 \phi^2/n}) \\
(iv) \quad & \|m' \mathcal{M}_{\hat{I}} H/\sqrt{n}\| = O_P(\sqrt{\phi/n} \sqrt{n} K^{-\alpha} + \sqrt{\phi^2/n}) \\
(v) \quad & \|W' \mathcal{M}_{\hat{I}} H/\sqrt{n}\| = O_P(\zeta_0(K) \sqrt{K/n} \sqrt{\phi/n} + \sqrt{K \phi^2/n} + K^{-\alpha} \sqrt{K \phi/s}) \\
(vi) \quad & \|W' \mathcal{P}_{\hat{I}} \epsilon/\sqrt{n}\| = O_P(\sqrt{K \phi^2/n}) \\
(vii) \quad & \|m' \mathcal{M}_{\hat{I}} \epsilon/\sqrt{n}\| = O_P(\sqrt{K^2 \phi^2/n})
\end{aligned}$$

*Proof.* Bounds for statement (i):

$$\begin{aligned}
\|W' \mathcal{P}_{\hat{I}} W/n\|^2 &= \sum_{k,l \leq K} (W'_k \mathcal{P}_{\hat{I}} W_l/n)^2 = \\
&= \sum_{k,l \leq K} (b(W_k; \hat{I})' Q W_l/n)^2 \leq \sum_{k,l \leq K} \|b(W_k; \hat{I})/\sqrt{n}\|_1^2 \|Q' W_l/\sqrt{n}\|_\infty^2 \\
&= \left( \sum_{k \leq K} \|b(W_k; \hat{I}_k)\|_1^2/n \right) \left( \sum_{l \leq K} \|Q' W_l/\sqrt{n}\|_\infty^2 \right) \\
&= O_P(K^2 s \phi/n^2) O_P(\phi/s) = O_P(K^2 \phi^2/n^2)
\end{aligned}$$

Where the last probability bounds follow from Lemma 3.

Next, bounds for statement (ii):

$$\begin{aligned}
\|m' \mathcal{M}_{\hat{I}} m/n\|^2 &= \sum_{k,l \leq K} (m'_k \mathcal{M}_{\hat{I}} m_l/n)^2 \leq \sum_{k,l \leq K} \|\mathcal{M}_{\hat{I}} m_k/\sqrt{n}\|^2 \|\mathcal{M}_{\hat{I}} m_l/\sqrt{n}\|^2 \\
&= \left( \sum_{k \leq K} \|\mathcal{M}_{\hat{I}} m_k/\sqrt{n}\|^2 \right)^2 = O_P(K \phi/n)^2
\end{aligned}$$

where again the final probability bounds follow from Lemma 3. This implies that  $\|m' \mathcal{M}_{\hat{I}} m/n\| = o_P(1)$  by  $K \phi/n \rightarrow 0$ .

Statement (iii): Let  $R_m = m - Q\Gamma$  track approximation errors in the first stage. Then

$$\begin{aligned}
\|m' \mathcal{M}_{\hat{I}} W/n\| &= \|m' W/n - m' \mathcal{P}_{\hat{I}} W/n\| \\
&= \|\Gamma Q' W/n + (m' - \Gamma' Q') W/n - m' \mathcal{P}_{\hat{I}} W/n\| \\
&= \|R'_m W/n + (\Gamma - \Gamma(\hat{I}))' Q' W/n\| \\
&\leq \|R'_m W/n\| + \|(\Gamma - \Gamma(\hat{I}))' Q' W/n\|
\end{aligned}$$

Then the first term in the last line is bounded by  $\|R'_m W/n\| = O_P(\zeta_0(K)\sqrt{K/n}\sqrt{\phi/n})$  while the second term has

$$\begin{aligned}
\|(\Gamma - \Gamma(\hat{I}_k))' Q' W/n\|^2 &= \sum_{k,l} [(\Gamma_k - \Gamma_k(\hat{I}_k))' Q' W_l/n]^2 \\
&= \sum_{k,l} \|\Gamma_k - \Gamma_k(\hat{I}_k)\|_1^2 \|Q' W_l/\sqrt{n}\|_\infty^2 \\
&= \left( \sum_k \|\Gamma_k - \Gamma_k(\hat{I}_k)\|_1^2 \right) \left( \sum_l \|Q' W_l/\sqrt{n}\|_\infty^2 \right) \\
&\leq \left( |\hat{I}_k| \sum_k \|\Gamma_k - \Gamma_k(\hat{I})\|^2 \right) \left( \sum_l \|Q' W_l/\sqrt{n}\|_\infty^2 \right) \\
&= O_P(Ks) O_P(\phi/n) K O_P(\phi/s)
\end{aligned}$$

With the last asertion following from Lemma 3. This gives Statement (iii) and  $\|m' \mathcal{M}_{\hat{I}} W/n\| = o_P(1)$ .

Statement (iv):

$$\begin{aligned}
\|F A' \hat{\Omega}^{-1} m' \mathcal{M}_{\hat{I}} G_2/\sqrt{n}\| &\leq \|F A' \hat{\Omega}^{-1}\| \max_{k \leq K} \|m' \mathcal{M}_{\hat{I}}/\sqrt{n}\| \sqrt{n} \|\mathcal{M}_{\hat{I}} H/\sqrt{n}\| \\
&= O_P(1) O_P(\sqrt{\phi/n}) \sqrt{n} O_P(\sqrt{\phi/n} + K^{-\alpha}) = o_P(1).
\end{aligned}$$

Statement (v):

$$\begin{aligned}
\|W' \mathcal{M}_{\hat{I}} H/\sqrt{n}\| &\leq \|(H - Q'\eta)' W/\sqrt{n}\| + \|(b(H; \hat{I}) - \eta)' Q' W/\sqrt{n}\| \\
&\leq O_P(\zeta_0 \sqrt{K/n} \sqrt{\phi/n}) + \sqrt{K} \max_{k \leq K} \|b(H; \hat{I}) - \eta\|_1 \|Q' W_k/\sqrt{n}\|_\infty \\
&\leq O_P(\zeta_0 \sqrt{K\phi/n}) + \sqrt{K} O_P(\sqrt{\phi/n} + K^{-\alpha}) O_P(\sqrt{\phi/s})
\end{aligned}$$

Statement (vi):

$$\begin{aligned}
\|W' \mathcal{P}_{\hat{I}} \epsilon/\sqrt{n}\| &= \|b(W; \hat{I}) Q' \epsilon/\sqrt{n}\| \\
&\leq \sqrt{K} \max_k \|b(W_k; \hat{I})\|_1 \|Q' \epsilon/\sqrt{n}\|_\infty \\
&= \sqrt{K} O_P(\sqrt{s\phi/n}) O_P(\sqrt{\phi/s})
\end{aligned}$$

Statement (vii): By reasoning similar to that for Lemma 4(iii), it is sufficient to bounds  $\|R'_m \epsilon/\sqrt{n}\| = O_P(\sqrt{\phi/n})$  and



$$\begin{aligned}
\|\Gamma(\widehat{I}) - \Gamma)'Q'\epsilon/\sqrt{n}\| &\leq \sqrt{K} \max_{k \leq K} |\Gamma(\widehat{I}) - \Gamma)'Q'\epsilon/\sqrt{n}| \\
&\leq \sqrt{K} \max_{k \leq K} \|\Gamma(\widehat{I}) - \Gamma\|_1 \|Q'\epsilon/\sqrt{n}\|_\infty \\
&\leq \sqrt{K} \max_k \sqrt{|\widehat{I}| + Ks} \|\Gamma(\widehat{I}) - \Gamma\| \|Q'\epsilon/\sqrt{n}\|_\infty \\
&= O_P(K\sqrt{\phi/n}) O_P(\sqrt{\phi/s}) = O_P(\sqrt{K^2\phi^2/n})
\end{aligned}$$

Then Statement (vii) follows and the proof of Lemma 3 is complete.  $\square$

**Lemma 5.**  $\max_{i \leq n} |h(z_i) - \widehat{h}(z_i)| = o_P(1)$

*Proof.* Let  $\widehat{\mathcal{T}} = \widehat{I} \cup \text{supp}(\eta)$ . Then  $\max_i |h(z_i) - \widehat{h}(z_i)| \leq \max_i |h(z_i) - q(z_i)'\eta| + \max_i |\widehat{h}(z_i) - q(z_i)'\eta|$ . The first term has the bound  $\max_i |h(x_i) - q(x_i)'\eta| = O_P(\sqrt{\phi/n})$  by assumption. A bound on the second term is obtained by the following:

$$\begin{aligned}
\max_i |\widehat{h}(z_i) - q(x_i)'\eta|^2 &= \max_i |q(x_i)'(\widehat{\eta} - \eta)|^2 \\
&\leq \max_i \|q_{\widehat{\mathcal{T}}}(z_i)\|^2 \|\widehat{\eta} - \eta\|^2 \\
&\leq |\widehat{\mathcal{T}}| \max_i \max_{j \leq L} |q_j(z_i)|^2 \|\widehat{\eta} - \eta\|^2 \\
&\leq O_P(Ks) \max_i \max_{j \leq L} |q_j(z_i)|^2 \|\widehat{\eta} - \eta\|^2
\end{aligned}$$

Then

$$\begin{aligned}
\|\widehat{\eta} - \eta\| &= \|b(y - \widehat{G}; \widehat{I}) - \eta\| = \|b(G; \widehat{I}) + b(H; \widehat{I}) + b(\epsilon; \widehat{I}) - b(\widehat{G}; \widehat{I}) - \eta\| \\
&\leq \|b(H; \widehat{I}) - \eta\| + \|b(\epsilon; \widehat{I})\| + \|b(G - \widehat{G}; \widehat{I})\|
\end{aligned}$$

First note that  $\|b(H; \widehat{I}) - \eta\| = O_P(\sqrt{\phi/n} + K^{-\alpha})$  by Lemma 3. Next,  $\|b(\epsilon; \widehat{I})\| \leq \sqrt{|\widehat{I}|} \phi_{\min}(\widehat{I}) \|Q'\epsilon/n\|_\infty = O_P(\sqrt{Ks}) O_P(1) \|Q'\epsilon/\sqrt{n}\|_\infty / \sqrt{n} = O_P(\sqrt{K\phi/n})$ . Finally,

$\|b(\widehat{G} - G; \widehat{I})\| \leq \|b(P(\beta - \widehat{\beta}); \widehat{I})\| + \|b(G - P\beta; \widehat{I})\|$ . The right term is  $\|b(G - P\beta; \widehat{I})\| = O_P(K^{-\alpha})$ . The left term is bounded by

$$\begin{aligned}
\|b(P(\beta - \widehat{\beta}); \widehat{I})\| &\leq \varphi_{\min}(|\widehat{T}|)^{-1} \sqrt{\widehat{T}} \max_j \left| \sum_i q_j(z_i) p(x_i)'(\beta - \widehat{\beta})/n \right| \\
&\leq \varphi_{\min}(|\widehat{T}|)^{-1} \sqrt{\widehat{T}} \max_j \sum_i |q_j(z_i)| |p(x_i)/n| \|(\beta - \widehat{\beta})\| \\
&= O_P(1) O_P(\sqrt{Ks}) \zeta_0(K) O_P(\sqrt{K}/n + K^{-\alpha}) \max_j \sum_i |q_j(z_i)|/n \\
&= o_P(1)
\end{aligned}$$

by the rate condition in Assumption 9. This completes the argument that  $\max_i |\widehat{h}(z_i) - q(z_i)' \eta| = o_P(1)$  and Lemma 5 follows.  $\square$

## APPENDIX D. TABLES

TABLE 1. Additively Separable Simulation Results: Low Dimensional Design, Average Derivative

	$n = 500$			$n = 1000$		
	Med. Bias	MAD	RP 5%	Med. Bias	MAD	RP 5%
A. High First Stage Signal/Noise, High Structural Signal/Noise						
Post-Double	-0.043	0.074	0.068	-0.028	0.047	0.074
Post-Double Set	-0.039	0.072	0.078	-0.030	0.049	0.088
Post-Double Ext	-0.049	0.077	0.076	-0.030	0.048	0.076
Post-Double Set+Ext	-0.048	0.074	0.086	-0.030	0.048	0.088
Post-Single I	-0.109	0.110	0.216	-0.095	0.096	0.308
Post-Single II	-0.116	0.117	0.236	-0.118	0.118	0.426
Series I	-0.009	0.072	0.076	-0.011	0.044	0.056
Series II	-0.020	0.078	0.100	-0.020	0.056	0.088
Oracle	0.001	0.059	0.058	-0.001	0.043	0.054
B. High First Stage Signal/Noise, Low Structural Signal/Noise						
Post-Double	-0.027	0.039	0.086	-0.017	0.025	0.074
Post-Double Set	-0.025	0.037	0.096	-0.017	0.026	0.078
Post-Double Ext	-0.028	0.039	0.092	-0.020	0.027	0.088
Post-Double Set+Ext	-0.030	0.040	0.102	-0.021	0.027	0.096
Post-Single I	-0.030	0.039	0.092	-0.024	0.029	0.108
Post-Single II	-0.033	0.041	0.102	-0.034	0.037	0.156
Series I	-0.002	0.036	0.072	-0.003	0.022	0.060
Series II	-0.008	0.040	0.090	-0.008	0.026	0.066
Oracle	-0.001	0.032	0.060	-0.002	0.022	0.048
C. Low First Stage Signal/Noise, High Structural Signal/Noise						
Post-Double	-0.034	0.042	0.104	-0.027	0.032	0.120
Post-Double Set	-0.034	0.040	0.108	-0.028	0.032	0.150
Post-Double Ext	-0.038	0.044	0.108	-0.028	0.033	0.124
Post-Double Set+Ext	-0.037	0.042	0.120	-0.028	0.032	0.156
Post-Single I	-0.056	0.058	0.244	-0.035	0.038	0.182
Post-Single II	-0.116	0.116	0.670	-0.117	0.117	0.918
Series I	-0.009	0.037	0.078	-0.011	0.023	0.060
Series II	-0.018	0.041	0.096	-0.018	0.030	0.094
Oracle	0.001	0.030	0.060	-0.000	0.022	0.052
D. Low First Stage Signal/Noise, Low Structural Signal/Noise						
Post-Double	-0.015	0.020	0.104	-0.008	0.013	0.078
Post-Double Set	-0.015	0.020	0.118	-0.008	0.013	0.090
Post-Double Ext	-0.015	0.021	0.114	-0.008	0.013	0.078
Post-Double Set+Ext	-0.015	0.021	0.126	-0.009	0.013	0.098
Post-Single I	-0.015	0.021	0.108	-0.009	0.014	0.082
Post-Single II	-0.032	0.033	0.260	-0.033	0.033	0.468
Series I	-0.003	0.018	0.074	-0.003	0.012	0.066
Series II	-0.007	0.021	0.098	-0.006	0.013	0.092
Oracle	-0.001	0.016	0.058	-0.000	0.011	0.050

Note: Results are based on 500 simulation replications. The table reports median bias (Med. Bias), median absolute deviation (MAD) and rejection frequency for a 5% level test (RP 5%) for nine different estimators of the function evaluated at the mean: the Post-Double proposed in this paper; the three variants of Post-Double (Post-Double Set, Post-Double Ext, Post-Double-Set+Ext) discussed in Sections 6,7; a post-model selection estimator (Post-Single I) based on selecting terms with Lasso on the reduced form equation only; a post-model selection estimator (Post-Single II) based on selecting terms using Lasso on the outcome equation (Post-Single II); and two estimators that uses a relatively small number of series terms (Series I, Series II); and an infeasible estimator that is explicitly given the control function.

TABLE 2. Additively Separable Models Simulation Results:  
Low Dimensional Design, Evaluation at the Mean

	$n = 500$			$n = 1000$		
	Med. Bias	MAD	RP 5%	Med. Bias	MAD	RP 5%
A. High First Stage Signal/Noise, High Structural Signal/Noise						
Post-Double	-0.064	0.182	0.074	-0.035	0.124	0.044
Post-Double Set	-0.082	0.152	0.074	-0.037	0.111	0.038
Post-Double Ext	-0.075	0.181	0.078	-0.040	0.124	0.044
Post-Double Set+Ext	-0.094	0.154	0.080	-0.041	0.110	0.038
Post-Single I	-0.177	0.218	0.128	-0.136	0.163	0.088
Post-Single II	-0.188	0.232	0.130	-0.175	0.189	0.118
Series I	-0.018	0.193	0.070	-0.004	0.122	0.050
Series II	-0.036	0.212	0.094	-0.027	0.137	0.060
Oracle	-0.010	0.181	0.058	0.011	0.118	0.040
B. High First Stage Signal/Noise, Low Structural Signal/Noise						
Post-Double	-0.066	0.200	0.070	-0.052	0.122	0.048
Post-Double Set	-0.168	0.189	0.130	-0.065	0.111	0.054
Post-Double Ext	-0.082	0.202	0.074	-0.055	0.123	0.056
Post-Double Set+Ext	-0.171	0.188	0.144	-0.074	0.115	0.070
Post-Single I	-0.083	0.209	0.072	-0.063	0.130	0.054
Post-Single II	-0.098	0.212	0.078	-0.093	0.138	0.058
Series I	-0.012	0.193	0.066	0.002	0.122	0.048
Series II	-0.021	0.212	0.094	-0.017	0.144	0.088
Oracle	0.005	0.182	0.074	-0.005	0.119	0.050
C. Low First Stage Signal/Noise, High Structural Signal/Noise						
Post-Double	-0.049	0.097	0.096	-0.036	0.068	0.060
Post-Double Set	-0.067	0.082	0.104	-0.040	0.060	0.068
Post-Double Ext	-0.052	0.097	0.098	-0.037	0.069	0.058
Post-Double Set+Ext	-0.070	0.083	0.114	-0.042	0.060	0.068
Post-Single I	-0.087	0.116	0.120	-0.050	0.074	0.074
Post-Single II	-0.187	0.189	0.286	-0.181	0.181	0.414
Series I	-0.017	0.097	0.078	-0.010	0.064	0.042
Series II	-0.032	0.107	0.098	-0.027	0.072	0.056
Oracle	-0.004	0.089	0.058	0.005	0.060	0.040
D. Low First Stage Signal/Noise, Low Structural Signal/Noise						
Post-Double	-0.045	0.101	0.082	-0.020	0.064	0.048
Post-Double Set	-0.115	0.122	0.224	-0.042	0.063	0.072
Post-Double Ext	-0.047	0.102	0.082	-0.023	0.063	0.052
Post-Double Set+Ext	-0.120	0.123	0.232	-0.042	0.064	0.078
Post-Single I	-0.046	0.098	0.084	-0.025	0.064	0.054
Post-Single II	-0.099	0.122	0.128	-0.090	0.099	0.112
Series I	-0.017	0.094	0.070	-0.005	0.062	0.054
Series II	-0.022	0.110	0.104	-0.014	0.074	0.086
Oracle	-0.007	0.090	0.070	-0.002	0.058	0.046

Note: Results are based on 500 simulation replications. The table reports median bias (Med. Bias), median absolute deviation (MAD) and rejection frequency for a 5% level test (RP 5%) for nine different estimators of the function evaluated at the mean: the Post-Double proposed in this paper; the three variants of Post-Double (Post-Double Set, Post-Double Ext, Post-Double-Set+Ext) discussed in Sections 6,7; a post-model selection estimator (Post-Single I) based on selecting terms with Lasso on the reduced form equation only; a post-model selection estimator (Post-Single II) based on selecting terms using Lasso on the outcome equation (Post-Single II); and two estimators that uses a relatively small number of series terms (Series I, Series II); and an infeasible estimator that is explicitly given the control function.

TABLE 3. Additively Separable Simulation Results: High Dimensional Design, Average Derivative

	$n = 500$			$n = 1000$		
	Med. Bias	MAD	RP 5%	Med. Bias	MAD	RP 5%
A. High First Stage Signal/Noise, High Structural Signal/Noise						
Post-Double	-0.010	0.060	0.044	-0.006	0.042	0.048
Post-Double Set	-0.008	0.060	0.042	-0.005	0.042	0.048
Post-Double Ext	-0.013	0.060	0.042	-0.006	0.043	0.052
Post-Double Set+Ext	-0.011	0.060	0.042	-0.005	0.043	0.046
Post-Single I	-0.056	0.071	0.096	-0.028	0.050	0.092
Post-Single II	-0.688	0.688	1.000	-0.692	0.692	1.000
Series I	-0.008	0.146	0.386	-0.027	0.111	0.424
Series II	-0.419	0.424	0.782	-0.458	0.458	0.872
Oracle	0.002	0.033	0.042	-0.001	0.025	0.052
B. High First Stage Signal/Noise, Low Structural Signal/Noise						
Post-Double	-0.007	0.031	0.040	-0.008	0.026	0.136
Post-Double Set	-0.007	0.030	0.052	-0.003	0.021	0.050
Post-Double Ext	-0.007	0.030	0.040	-0.009	0.026	0.136
Post-Double Set+Ext	-0.008	0.029	0.056	-0.003	0.022	0.052
Post-Single I	-0.014	0.032	0.056	-0.011	0.027	0.140
Post-Single II	-0.357	0.357	1.000	-0.356	0.356	1.000
Series I	-0.002	0.072	0.390	-0.291	0.291	0.924
Series II	-0.155	0.168	0.682	-0.313	0.313	0.988
Oracle	-0.001	0.025	0.040	-0.001	0.020	0.058
C. Low First Stage Signal/Noise, High Structural Signal/Noise						
Post-Double	-0.010	0.031	0.056	-0.005	0.021	0.050
Post-Double Set	-0.008	0.030	0.054	-0.003	0.021	0.052
Post-Double Ext	-0.011	0.030	0.062	-0.006	0.022	0.052
Post-Double Set+Ext	-0.010	0.030	0.056	-0.004	0.021	0.062
Post-Single I	-0.017	0.032	0.068	-0.009	0.022	0.062
Post-Single II	-0.689	0.689	1.000	-0.692	0.692	1.000
Series I	-0.004	0.073	0.386	-0.022	0.060	0.456
Series II	-0.414	0.414	0.858	-0.449	0.449	0.916
Oracle	0.000	0.017	0.044	-0.001	0.013	0.054
D. Low First Stage Signal/Noise, Low Structural Signal/Noise						
Post-Double	-0.004	0.015	0.044	-0.005	0.014	0.136
Post-Double Set	-0.004	0.015	0.052	-0.001	0.011	0.066
Post-Double Ext	-0.005	0.016	0.044	-0.005	0.013	0.138
Post-Double Set+Ext	-0.005	0.016	0.052	-0.002	0.011	0.064
Post-Single I	-0.005	0.016	0.044	-0.005	0.013	0.142
Post-Single II	-0.357	0.357	1.000	-0.356	0.356	1.000
Series I	-0.001	0.036	0.398	-0.295	0.295	0.986
Series II	-0.153	0.153	0.808	-0.317	0.317	0.994
Oracle	-0.001	0.012	0.048	-0.000	0.010	0.078

Note: Results are based on 500 simulation replications. The table reports median bias (Med. Bias), median absolute deviation (MAD) and rejection frequency for a 5% level test (RP 5%) for nine different estimators of the function evaluated at the mean: the Post-Double proposed in this paper; the three variants of Post-Double (Post-Double Set, Post-Double Ext, Post-Double-Set+Ext) discussed in Sections 6,7; a post-model selection estimator (Post-Single I) based on selecting terms with Lasso on the reduced form equation only; a post-model selection estimator (Post-Single II) based on selecting terms using Lasso on the outcome equation (Post-Single II); and two estimators that uses a relatively small number of series terms (Series I, Series II); and an infeasible estimator that is explicitly given the control function.

TABLE 4. Additively Separable Simulation Results: HighDimensional Design, Evaluation at the Mean

	$n = 500$			$n = 1000$		
	Med. Bias	MAD	RP 5%	Med. Bias	MAD	RP 5%
A. High First Stage Signal/Noise, High Structural Signal/Noise						
Post-Double	-0.031	0.214	0.034	0.006	0.151	0.038
Post-Double Set	-0.102	0.182	0.066	-0.016	0.147	0.054
Post-Double Ext	-0.038	0.220	0.036	0.002	0.154	0.038
Post-Double Set+Ext	-0.113	0.185	0.066	-0.021	0.148	0.052
Post-Single I	-0.126	0.232	0.100	-0.051	0.148	0.054
Post-Single II	-1.654	1.654	1.000	-1.662	1.662	1.000
Series I	0.066	0.546	0.440	-0.091	0.367	0.376
Series II	-0.964	0.979	0.706	-1.030	1.030	0.812
Oracle	0.014	0.192	0.054	0.015	0.131	0.054
B. High First Stage Signal/Noise, Low Structural Signal/Noise						
Post-Double	-0.063	0.190	0.042	-0.032	0.158	0.136
Post-Double Set	-0.137	0.175	0.108	-0.048	0.124	0.062
Post-Double Ext	-0.066	0.190	0.044	-0.035	0.155	0.136
Post-Double Set+Ext	-0.139	0.175	0.118	-0.049	0.125	0.062
Post-Single I	-0.088	0.191	0.044	-0.047	0.160	0.136
Post-Single II	-1.207	1.207	0.982	-1.187	1.187	1.000
Series I	0.007	0.482	0.442	-0.753	0.753	0.294
Series II	-0.481	0.619	0.534	-0.760	0.760	0.354
Oracle	-0.033	0.176	0.052	-0.000	0.135	0.100
C. Low First Stage Signal/Noise, High Structural Signal/Noise						
Post-Double	-0.029	0.116	0.046	-0.002	0.076	0.042
Post-Double Set	-0.088	0.111	0.114	-0.018	0.076	0.066
Post-Double Ext	-0.031	0.116	0.048	-0.004	0.076	0.042
Post-Double Set+Ext	-0.094	0.115	0.118	-0.021	0.076	0.060
Post-Single I	-0.050	0.116	0.054	-0.013	0.077	0.042
Post-Single II	-1.674	1.674	1.000	-1.660	1.660	1.000
Series I	0.026	0.275	0.442	-0.054	0.190	0.428
Series II	-0.937	0.938	0.842	-1.055	1.055	0.890
Oracle	-0.000	0.096	0.058	0.008	0.068	0.052
D. Low First Stage Signal/Noise, Low Structural Signal/Noise						
Post-Double	-0.045	0.099	0.046	-0.018	0.078	0.138
Post-Double Set	-0.125	0.128	0.252	-0.042	0.070	0.092
Post-Double Ext	-0.046	0.099	0.046	-0.018	0.078	0.140
Post-Double Set+Ext	-0.127	0.129	0.252	-0.044	0.069	0.088
Post-Single I	-0.046	0.098	0.046	-0.021	0.078	0.142
Post-Single II	-1.201	1.201	1.000	-1.196	1.196	1.000
Series I	-0.004	0.241	0.440	-0.752	0.752	0.836
Series II	-0.475	0.495	0.658	-0.758	0.758	0.890
Oracle	-0.024	0.088	0.052	0.004	0.070	0.098

Note: Results are based on 500 simulation replications. The table reports median bias (Med. Bias), median absolute deviation (MAD) and rejection frequency for a 5% level test (RP 5%) for nine different estimators of the function evaluated at the mean: the Post-Double proposed in this paper; the three variants of Post-Double (Post-Double Set, Post-Double Ext, Post-Double-Set+Ext) discussed in Sections 6,7; a post-model selection estimator (Post-Single I) based on selecting terms with Lasso on the reduced form equation only; a post-model selection estimator (Post-Single II) based on selecting terms using Lasso on the outcome equation (Post-Single II); and two estimators that uses a relatively small number of series terms (Series I, Series II); and an infeasible estimator that is explicitly given the control function.

TABLE 5. Average Effects of Export on HIV:  
UNAIDS Incidence

	Ave. Derivative	95% Confidence Interval
	1. Log Export Value (WDI)	
Baseline:	0.015	[ -0.135 0.165 ]
All:	0.044	[ -0.032 0.121 ]
Post-Double:	0.013	[ -0.104 0.131 ]
	$n = 720, K = 5, L = 78$ Selected vars: lag-incidence, $t \times \text{incidence}_0$ , $\sin(2\pi t/4) \times \text{incidence}_0$ , $\cos(2\pi t/4) \times \text{gdp}_0$ , $\sin(2\pi t/8) \times \text{gdp}_0$ , $\text{incidence}_0 \times \text{lag-incidence}$	
	1. Log Export Value (WDI)	
Baseline:	0.008	[ -0.238 0.254 ]
All:	0.038	[ -0.052 0.128 ]
Post-Double:	0.038	[ -0.117 0.195 ]
	$n = 747, K = 5, L = 78$ Selected vars: lag-incidence, $t \times \text{incidence}_0$ , $\sin(2\pi t/4) \times \text{incidence}_0$ , $\cos(2\pi t/4) \times 1_{\text{region } 2}$ , $\sin(2\pi t/8) \times \text{incidence}_0$ , $\cos(2\pi t/8) \times 1_{\text{region } 2}$ , $\text{incidence}_0 \times \text{lag-incidence}$ , $\text{pop} \times \text{gdp}_0$	
	1. Log Export Value (WDI)	
Baseline:	0.028	[ -0.108 0.166 ]
All:	-0.041	[ -0.131 0.048 ]
Post-Double:	0.059	[ -0.053 0.172 ]
	$n = 747, K = 5, L = 78$ Selected vars: lag-incidence, $t \times \text{incidence}_0$ , $\sin(2\pi t/4) \times \text{incidence}_0$ , $\sin(2\pi t/8) \times 1_{\text{region } 2}$ , $\text{incidence}_0 \times \text{lag-incidence}$ , $\text{pop} \times \text{pop}_0$	

The table presents estimates of the sample weighted average derivative of growth in HIV incidence with respect to growth in export. The estimates are calculated using a baseline model, using post-double selection over the set of conditioning variables, and over the entire set of conditioning variables. All variables are generated from data given in [30].  $n$  gives the number of total observations (not observational units). Estimates are based on a  $K$ -order Hermite polynomial in the export measure.  $L$  gives the number of conditioning variables.



TABLE 6. Average effects of Export on HIV:  
Death-based Incidence

	Ave. Derivative	95% Confidence Interval
	1. Log Export Value (WDI)	
Baseline:	0.915	[ 0.294 1.536 ]
All:	0.239	[ -0.755 1.235 ]
Post-Double:	0.904	[ 0.348 1.461 ]
	$n = 161, K = 3, L = 78$ Selected vars: $\sin(2\pi t/8) \times x_0$	
	2. Log Export Value (NBER)	
Baseline:	0.733	[ 0.274 1.191 ]
All:	0.614	[ 0.330 0.899 ]
Post-Double:	0.646	[ 0.112 1.180 ]
	$n = 166, K = 3, L = 78$ Selected vars: $t^2 \times \text{export}_0, \cos(2\pi t/4) \times \text{pop}, \sin(2\pi t/8) \times \text{pop}_0, \cos(2\pi t/8) \times \text{pop}_0, \cos(2\pi t/8) \times \text{pop}$	
	3. Log Export Volume (NBER)	
Baseline:	0.405	[ -0.354 1.165 ]
All:	0.398	[ -0.015 0.811 ]
Post-Double:	0.366	[ -0.347 1.080 ]
	$n = 166, K = 3, L = 78$ Selected vars: $t^2 \times \text{export}_0, \cos(2\pi t/8) \times \text{pop}_0$	

The table presents estimates of the sample weighted average derivative of growth in HIV incidence with respect to growth in export. The estimates are calculated using a baseline model, using post-double selection over the set of conditioning variables, and over the entire set of conditioning variables. All variables are generated from data given in [30].  $n$  gives the number of total observations (not observational units). Estimates are based on a  $K$ -order Hermite polynomial in the export measure.  $L$  gives the number of conditioning variables.

## REFERENCES

- [1] J. Bai and S. Ng. Forecasting economic time series using targeted predictors. *Journal of Econometrics*, 146:304–317, 2008.
- [2] J. Bai and S. Ng. Boosting diffusion indices. *Journal of Applied Econometrics*, 24, 2009.
- [3] A. Belloni, D. Chen, V. Chernozhukov, and C. Hansen. Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica*, 80:2369–2429, 2012. Arxiv, 2010.
- [4] A. Belloni and V. Chernozhukov. Least squares after model selection in high-dimensional sparse models. *Bernoulli*, 19(2):521–547, 2013. ArXiv, 2009.
- [5] A. Belloni, V. Chernozhukov, and C. Hansen. Lasso methods for gaussian instrumental variables models. 2010 arXiv:[math.ST], <http://arxiv.org/abs/1012.1297>, 2010.
- [6] A. Belloni, V. Chernozhukov, and C. Hansen. Inference for high-dimensional sparse econometric models. *Advances in Economics and Econometrics. 10th World Congress of Econometric Society. August 2010*, III:245–295, 2013.
- [7] A. Belloni, V. Chernozhukov, C. Hansen, and D. Kozbur. Inference in high dimensional panel models with an application to gun control. *ArXiv:1411.6507*, 2014.
- [8] Alexandre Belloni, Victor Chernozhukov, Ivan Fernández-Val, and Christian Hansen. Program evaluation with high-dimensional data. *arXiv:1311.2645*, 2014.
- [9] Alexandre Belloni, Victor Chernozhukov, and Christian Hansen. Inference on treatment effects after selection amongst high-dimensional controls with an application to abortion on crime. *Review of Economic Studies*, 81(2):608–650, 2014.
- [10] P. J. Bickel, Y. Ritov, and A. B. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics*, 37(4):1705–1732, 2009.
- [11] P. Bühlmann and S. van de Geer. *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer, 2011.
- [12] F. Bunea, A. Tsybakov, and M. H. Wegkamp. Sparsity oracle inequalities for the lasso. *Electronic Journal of Statistics*, 1:169–194, 2007.
- [13] F. Bunea, A. B. Tsybakov, , and M. H. Wegkamp. Aggregation and sparsity via  $\ell_1$  penalized least squares. In *Proceedings of 19th Annual Conference on Learning Theory (COLT 2006) (G. Lugosi and H. U. Simon, eds.)*, pages 379–391, 2006.
- [14] F. Bunea, A. B. Tsybakov, and M. H. Wegkamp. Aggregation for Gaussian regression. *The Annals of Statistics*, 35(4):1674–1697, 2007.
- [15] E. Candès and T. Tao. The Dantzig selector: statistical estimation when  $p$  is much larger than  $n$ . *Ann. Statist.*, 35(6):2313–2351, 2007.
- [16] X. Chen. Large sample sieve estimation of semi-nonparametric models. *Handbook of Econometrics*, 6:5559–5632, 2007.
- [17] J. Fan and J. Lv. Sure independence screening for ultra-high dimensional feature space. *Journal of the Royal Statistical Society Series B*, 70(5):849–911, 2008.
- [18] Ildiko E. Frank and Jerome H. Friedman. A statistical view of some chemometrics regression tools. *Technometrics*, 35(2):109–135, 1993.
- [19] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York, NY, 2009.
- [20] Jian Huang, Joel L. Horowitz, and Fengrong Wei. Variable selection in nonparametric additive models. *Ann. Statist.*, 38(4):2282–2313, 2010.
- [21] Adel Javanmard and Andrea Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. *Journal of Machine Learning Research*, 15:2869–2909, 2014.
- [22] Bing-Yi Jing, Qi-Man Shao, and Qiying Wang. Self-normalized cramr-type large deviations for independent random variables. *Ann. Probab.*, 31(4):2167–2215, 2003.
- [23] Keith Knight. Shrinkage estimation for nearly singular designs. *Econometric Theory*, 24:323–337, 2008.

- [24] V. Koltchinskii. Sparsity in penalized empirical risk minimization. *Ann. Inst. H. Poincaré Probab. Statist.*, 45(1):7–57, 2009.
- [25] Hannes Leeb and Benedikt M. Pötscher. Can one estimate the unconditional distribution of post-model-selection estimators? *Econometric Theory*, 24(2):338–376, 2008.
- [26] K. Lounici. Sup-norm convergence rate and sign concentration property of lasso and dantzig estimators. *Electron. J. Statist.*, 2:90–102, 2008.
- [27] K. Lounici, M. Pontil, A. B. Tsybakov, and S. van de Geer. Taking advantage of sparsity in multi-task learning. *arXiv:0903.1468v1 [stat.ML]*, 2010.
- [28] N. Meinshausen and B. Yu. Lasso-type recovery of sparse representations for high-dimensional data. *Annals of Statistics*, 37(1):2246–2270, 2009.
- [29] Whitney K. Newey. Convergence rates and asymptotic normality for series estimators. *Journal of Econometrics*, 79:147–168, 1997.
- [30] Emily Oster. Routes of infection: Exports and hiv incidence in sub-saharan africa. *Journal of the European Economic Association*, 10(5):1025–1058, 2012.
- [31] Benedikt M. Pötscher. Confidence sets based on sparse estimators are necessarily large. *Sankhyā*, 71(1, Ser. A):1–18, 2009.
- [32] P. M. Robinson. Root- $N$ -consistent semiparametric regression. *Econometrica*, 56(4):931–954, 1988.
- [33] M. Rosenbaum and A. B. Tsybakov. Sparse recovery under matrix uncertainty. *arXiv:0812.2818v1 [math.ST]*, 2008.
- [34] M. Rudelson and S. Zhou. Reconstruction from anisotropic random measurements. *ArXiv:1106.1151*, 2011.
- [35] Mark Rudelson and Roman Vershynin. On sparse reconstruction from fourier and gaussian measurements. *Communications on Pure and Applied Mathematics*, 61:10251045, 2008.
- [36] Charles J. Stone. Additive regression and other nonparametric models. *The Annals of Statistics*, 13(2):689–705, 1985.
- [37] R. Tibshirani. Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B*, 58:267–288, 1996.
- [38] S. A. van de Geer. High-dimensional generalized linear models and the lasso. *Annals of Statistics*, 36(2):614–645, 2008.
- [39] Sara van de Geer, Peter Bhlmann, Yaacov Ritov, and Ruben Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann. Statist.*, 42(3):1166–1202, 06 2014.
- [40] M. Wainwright. Sharp thresholds for noisy and high-dimensional recovery of sparsity using  $\ell_1$ -constrained quadratic programming (lasso). *IEEE Transactions on Information Theory*, 55:2183–2202, May 2009.
- [41] C.-H. Zhang and J. Huang. The sparsity and bias of the lasso selection in high-dimensional linear regression. *Ann. Statist.*, 36(4):1567–1594, 2008.
- [42] Cun-Hui Zhang and Stephanie S. Zhang. Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):217–242, 2014.
- [43] S. Zhou. Restricted eigenvalue conditions on subgaussian matrices. *ArXiv:0904.4723v2*, 2009.